AI for BIM-Based Sustainable Building Design

integrating knowledge discovery and semantic data modelling for evidence-based design decision support

Petrova, Ekaterina Aleksandrova

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

AI FOR BIM-BASED SUSTAINABLE BUILDING DESIGN

INTEGRATING KNOWLEDGE DISCOVERY AND SEMANTIC DATA MODELLING FOR EVIDENCE-BASED DESIGN DECISION SUPPORT

by

Ekaterina Aleksandrova Petrova

AALBORG UNIVERSITY
DENMARK

Dissertation submitted 2019

Dissertation submitted:	June 2019
PhD supervisor:	Associate Prof. Kjeld Svidt, Aalborg University
Assistant PhD supervisor:	Associate Prof. Rasmus Lund Jensen, Aalborg University
PhD committee:	Prof. Per Heiselberg, Aalborg University (chairman) Prof. Bauke De Vries, Eindhoven University of Technology Prof. Jakob Beetz, RWTH Aachen University
PhD Series:	Faculty of Engineering and Science, Aalborg University

ISSN: xxxx- xxxx
ISBN: xxx-xx-xxxx-xxx-x

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2019

CURRICULUM VITAE

Personal Information

Name: Ekaterina Aleksandrova Petrova

Date of birth: 13.12.1985

Nationality: Bulgarian

E-mail: ep@civil.aau.dk

Phone: +45 23 92 54 25

Education

Feb 2014- Jan 2016	MSc in Technology in Management in the Building Industry, Aalborg University, Aalborg, Denmark
Aug 2008- Feb 2012	BSc in Architectural Technology and Construction Management, University College of Northern Denmark, Aalborg, Denmark

Professional Experience

Jun 2016- May 2019	Ph.D. Candidate, Department of Civil Engineering, Aalborg University, Aalborg, Denmark
Mar 2018- Jun 2018	Guest Researcher and Teaching Assistant, Ghent University, Ghent, Belgium
Mar 2016- May 2016	Research Assistant, Department of Civil Engineering, Aalborg University, Aalborg, Denmark
Aug 2012- Jan 2013	Lecturer, University College of Northern Denmark, Aalborg, Denmark

Honors and Awards

April 2018	Professor P. Ole Fanger's Research Grant
September 2017	Education and Research in Computer Aided Architectural Design in Europe, Young Researcher Grant

Research Areas

BIM, Knowledge Discovery in Databases, Semantic Data Modelling, Sustainable Design, Decision Support Systems

Teaching experience

Knowledge Management in Architecture, Engineering and Construction (course coordinator)
Introduction to Building Information Management (course coordinator)

ENGLISH SUMMARY

Sustainable building design requires an interplay between multidisciplinary input and fulfilment of various criteria, which need to align into one high-performing whole: the building. Building Information Modelling has already brought a profound change in Architecture, Engineering and Construction by enabling efficient collaborative workflows. Combined with the power of statistical and symbolic Artificial Intelligence approaches (e.g. machine learning, semantic query techniques, inference machines, etc.) and the richness of data, these technologies can foster accurate prediction of design outcomes and help uncover valuable hidden knowledge in the performance of the existing built environment. Such knowledge has the potential to create an information cycle that can redefine building design and serve as valuable evidence for design decision support.

However, despite the technological advancements, the gap between designed and measured building performance remains. Design decision-making still, to a large extent, relies on rules of thumb and previous experiences, and not on sound evidence. Consequently, design practice is neither sufficiently data-driven nor evidence-based. Performance mismatches also occur due to inaccurate predictions and assumptions, lack of data integration and sharing across domains, poor modelling and collaboration, etc. Research has investigated possible solutions to eliminate these causes, but few attempts have been made to sever the problem at its core- the lack of feedback loop from building operation to design. In response to the latter, this research effort attempts to unlock the potential of Artificial Intelligence approaches to establish the missing feedback loop and enhance human decision-making capabilities.

Therefore, this thesis aims to demonstrate how knowledge discovery, representation and retrieval techniques can be integrated to create the missing link between building operation and design and inform sustainable BIM-based design decision-making in an evidence-based, context-aware and user-centred way.

To achieve the research objective, the thesis presents an in-depth analysis of the diverse building data sources and types and outlines how the data can be analysed to discover valuable knowledge. Based on the results of that analysis and an extensive literature review, a framework for performance-oriented design decision support relying on BIM, data mining and semantic data modelling is proposed. Furthermore, motif discovery and association rule mining are performed on operational building data from two use case buildings to uncover performance insights. The discovered knowledge is then represented in an ecosystem of (semantic) data to create a knowledge base enriched with building performance patterns. A significant challenge, namely the interpretation of the discovered knowledge, is approached using linked data and crowdsourcing techniques, which results in contextualised networks of building data and knowledge annotated by human domain experts. Finally, the thesis

demonstrates how the created knowledge ecosystem can reach the building design professionals through evidence-based recommendations based on semantic relatedness between concepts and determined by the users' profile and context.

As such, the presented future-proof holistic technological approach enables a robust user-centred mechanism that allows knowledge discovery, representation, contextualisation and reuse and achieves the targeted, evidence-based decision support in BIM-based sustainable design processes.

DANSK RESUME

Bæredygtigt bygningsdesign kræver et samspil mellem tværfaglige input og opfyldelse af forskellige kriterier, som skal tilpasses og føre til en højt performende helhed: bygningen. Building Information Modeling (BIM) har allerede bibragt markante ændringer i denne retning ved at muliggøre effektive samarbejdsprocesser. Kombineret med metoderne fra statistisk og symbolsk kunstig intelligens (eks. maskinlæring, semantiske forespørgselsmetoder, inferens-maskiner) samt rige data fra byggeindustrien, gør disse teknologier i stand til at fremme præcise forudsigelser af designresultater. Yderligere kan de medvirke til at afdække værdifuld uopdaget viden, ved udførelsen af de eksisterende bygninger.

Denne viden har potentiale til at skabe en informationscyklus, der kan omdefinere en bygnings designproces og tjene som værdifuld evidens for designbeslutningsstøtte. På trods af disse teknologiske fremskridt ser vi desværre stadig stor forskel mellem bygningers beregnede performance og det, som kan måles under praktiske driftsforhold. Designbeslutningstagning er i høj grad baseret på tommelfingerregler og tidligere erfaringer, frem for håndfaste beviser. Derfor er designpraksis hverken tilstrækkeligt datadrevet eller evidensbaseret. Forskellen mellem beregnet og målt performance opstår også grundet unøjagtige forudsigelser og antagelser, manglende dataintegration samt manglende deling på tværs af fagdiscipliner, mangelfuld modellering og samarbejde mv.

Videnskaben har undersøgt mulige løsningsmodeller, der kan eliminere disse årsager, men kun få forsøg er rapporteret, hvor problemet er forsøgt løst ved dets kerne; manglen på feedback loop fra bygningens driftsforhold til design. Som reaktion på sidstnævnte forsøger forskningsindsatsen, som er beskrevet i denne afhandling, at realisere potentialet fra kunstig intelligens tilgange ved at etablere det manglende feedback loop og forbedre den menneskelige beslutningstagning.

Denne afhandling søger at demonstrere, hvordan videnopdagelse, videnrepræsentation og hentningsteknikker kan integreres således, at den manglende forbindelse mellem bygningens driftsforhold og design understøttes. Endvidere søger afhandlingen at understøtte bæredygtig BIM-baseret beslutningstagning på en evidensbaseret, kontekstbevidst og brugercentreret måde. For at opnå disse forskningsmål præsenterer afhandlingen en grundig analyse af de forskellige byggedatakilder og -typer, og beskriver desuden, hvordan data kan analyseres for at opdage værdifuld viden.

Baseret på resultaterne af denne analyse og et omfattende litteraturstudie, foreslås en teoretisk ramme for performanceorienteret designbeslutningsstøtte, baseret på BIM, data mining og semantisk datamodellering. Desuden udføres mønstergenkendelse og associeringsregelmining på sensordata fra to bygninger i drift for at skabe indsigt i

performancesammenhænge. Den opdagede viden er dernæst repræsenteret i et økosystem af (semantiske) data for at skabe en vidensbase beriget med bygningens performancemønstre. En væsentlig udfordring, nemlig fortolkningen af den opdagede viden, er tilvejebragt ved hjælp af sammenkædede data og crowdsourcing teknikker. Dette resulterer i kontekstualiserede netværk til opbygning af data og viden annoteret af menneskelige domæneeksperter.

Endelig demonstrerer afhandlingen, hvordan dette videnøkosystem kan formidles til byggeprojektets fagfolk gennem evidensbaserede anbefalinger baseret på semantisk tilknytning mellem begreber og bestemt af brugerens profil og kontekst.

Den præsenterede fremtidssikrede holistiske teknologiske tilgang muliggør en robust brugercentreret mekanisme, der tillader videnopdagelse, repræsentation, kontekstualisering og genanvendelse og opnår målrettet, evidensbaseret beslutningsstøtte for bæredygtige BIM-baserede designprocesser.

PREFACE

The work presented in this thesis is a part of a PhD project funded by the Department of Civil Engineering, Aalborg University. The research has been carried out by Ekaterina Aleksandrova Petrova in the period from 1st of June 2016 to 31st of May 2019. The author greatly appreciates the opportunity provided by Aalborg University.

PAPER OVERVIEW

This thesis is paper-based and consists of the following collection of papers :

Paper A *"Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data"*
Petrova, E.; Pauwels, P.; Svidt, K.; Jensen, R.L.
Architectural Engineering and Design Management 2018
Special Issue on Intelligent Building Paradigms and Data-Driven Models of Innovation

Paper B *"In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data"*
Petrova, E.; Pauwels, P.; Svidt, K.; Jensen, R.L.
Advances in Informatics and Computing in Civil and Construction Engineering 2018

Paper C *"Data mining and semantics for decision support in sustainable BIM-based design"*
Petrova, E.; Pauwels, P.; Svidt, K.; Jensen, R.L.
Advanced Engineering Informatics, submitted April 2019

Paper D *"From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches"*
Petrova, E.; Pauwels, P.; Svidt, K.; Jensen, R.L.
12th European Conference on Product and Process Modelling 2018

Paper E *"Crowdsourcing building performance patterns for evidence-based decision support in sustainable building design"*
Petrova, E.; Pauwels, P.; Svidt, K.; Jensen, R.L.
Automation in Construction, submitted May 2019

Paper F *"Semantic data mining and linked data for a recommender system in the AEC industry"*
Petrova, E.; Pauwels, P.; Svidt, K.; Jensen, R.L.
European Conference on Computing in Construction 2019

The thesis consists of an extended summary and papers A-F, which have been included in the appendices. The purpose of the extended summary is to be able to represent the work as a whole and makes it possible to understand the research area, objectives and contributions entirely, without necessarily referring to the appendices. Papers A-F present the contributions in full detail and elaborate more extensively on the methodology and the results. Each chapter provides a reference to the paper(s) that it is based on. Unless otherwise stated, all illustrations are the author's own work.

Besides papers A-F, the author has also worked on one academic journal article, four conference papers and one industry journal article during or before the PhD study. These papers are not a part of this thesis and should, therefore, not be evaluated, but they are included to showcase the additional research activities the author has been involved in and the topics that have been investigated. The journal article is a result of a design studio research experiment with students directly related to the objectives of the thesis and the conference papers explore various aspects of information management in the built environment.

Paper G *"Pattern ReCognition in Sustainable Architectural Design: Assessing the Effects of Context and Team Dynamics with Protocol Studies"*
Petrova, E.; Pauwels, P.
Research in Engineering Design, submitted June 2018

Paper H *"Development of an Information Delivery Manual for Early Stage BIM-based Energy Performance Assessment and Code Compliance as a Part of DGNB Pre-Certification"*
Petrova, E.; Romanska, I.; Stamenov, M.; Svidt, K.; Jensen, R.L.
IBPSA Building Simulation 2017

Paper I *"Automation of Geometry Input for Building Code Compliance Check"*
Petrova, E.; Johansen, P.L.; Jensen, R.L.; Maagaard, S.; Svidt, K.
Joint Conference on Computing in Construction 2017

Paper J *"Integrating Virtual Reality and BIM for End-User Involvement in Building Design: a case study"*
Petrova, E.; Rasmussen, M.; Jensen, R.L.; Svidt, K.
Joint Conference on Computing in Construction 2017

Paper K *"Let the data tell you the truth. Data-driven decision support for high-performance building design"*
Petrova, E.
HVAC Magasinet 2018

ACKNOWLEDGEMENTS

I would like to extend my warmest and most sincere thanks to many people, in many countries, who generously contributed to the completion of this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisors, Associate Professor Kjeld Svidt and Associate Professor Rasmus Lund Jensen. Words cannot describe how thankful I am for your guidance, patience, motivation, and support throughout my PhD journey. You saw something in me that I myself did not know I had. By giving me the chance to pursue this career, you unknowingly gave me my biggest passion in life. Thank you for your faith in me. I could not have imagined having better supervisors and for that, I am forever grateful.

Besides my supervisors, I would also like to thank my main research collaborator, Dr. Pieter Pauwels. This thesis would not have been what it is without your mentorship, expertise, insightful comments, and hard questions. You saw my research vision right away and assured me that it is worth believing in. Thank you for the constructive, inspiring and challenging collaboration, and for sharing your immense knowledge with me.

I would also like to say thank you to the administrative staff at the Department of Civil Engineering, especially Pernille Bisgaard, Anja Bloch, Mette Vegeberg, Linda Andersen and Vivi Søndergaard. Sincere thanks to my current and former fellow PhD colleagues Lasse and Mai for the friendship, the insightful discussions, and for always being there when PhD life got tough- a PhD worry shared is a PhD worry halved. Special mention goes to Yovko, Kirstine, Holly, Peter and Anne for being the best people to share "the corner office" with and for making the last three years an even better experience.

A very special thank you to my dear friends and former classmates Iva and Martin. The master thesis that we did together and the topic that we chose laid the foundations of my research interests and made me want to pursue a research career. Thank you for trusting me and embarking on that journey with me. I would not have been here without you.

From the bottom of my heart, I would like to thank my parents, for being the best role models one could ever wish for and for always being there for me. Heartfelt thanks to my family-in-law, for the continuous help and encouragement. Last, but not least, to the two men in my life, Simon and Alexander, thank you for your almost unbelievable and unconditional support, and for always believing in me, even when I didn't. This accomplishment would not be possible without you. Thank you!

To my beloved parents, Diana and Aleksandar,

for raising me to believe that anything is possible,

and to Simon and Alexander,

for supporting me unconditionally and making this journey possible.

TABLE OF CONTENTS

NOMENCLATURE- ABBREVIATIONS

AEC	Architecture, Engineering and Construction
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
ARM	Association Rule Mining
BIM	Building Information Modelling
BOT	Building Topology Ontology
BMS	Building Monitoring Systems
CBR	Case-Based Reasoning
CDE	Common Data Environment
DL	Description Logics
DSS	Decision Support System
GUI	Graphical User Interface
HTM	Hierarchical Temporal Memory
HTTP	HyperText Transfer Protocol
ICT	Information and Communication Technology
IFC	Industry Foundation Classes
IRI	Internationalized Resource Identifier
KBC	Knowledge Base Construction
KBN	Knowledge-Based Neurocomputing
KBS	Knowledge-Based System
KDD	Knowledge Discovery in Databases
LBD	Linked Building Data
LDSD	Linked Data Semantic Distances

LRS	Longest Repeated Substring
NZEB	Nearly Zero-Energy Building
OWL	Web Ontology Language
PAA	Piecewise Approximate Aggregation
RDF	Resource Description Framework
RDFS	RDF Schema
RH	Relative Humidity
SAX	Symbolic Aggregate Approximation
SCADA	Supervisory Control and Data Acquisition
SPARQL	SPARQL protocol and RDF Query Language
SSN	Semantic Sensor Network
SQL	Structured Query Language
SWRL	Semantic Web Rule Language
TVOC	Total Volatile Organic Compounds
URI	Unique Resource Identifier
WKT	Well-Known Text

CHAPTER 1. INTRODUCTION

"The most dangerous phrase in the language is, "We've always done it this way."

Rear Admiral Grace Hopper

1.1. TOWARDS EVIDENCE-BASED SUSTAINABLE BUILDING DESIGN

Buildings account for about 40% of the total energy use in Europe and one third of the global CO_2 emissions. About 60% of the energy needs are attributed to indoor space heating and cooling, and water heating (International Energy Agency, 2013). In addition, people spend about 90% of their time indoors (Klepeis et al., 2001). These significant contributions have put the built environment amongst the main priorities in reaching critical energy and environmental performance objectives (The European Parliament and Council, 2010). Due to the complex relationship between building performance, climate change, resource depletion and occupant well-being, contemporary building design practices have been amended to integrate sustainability as a fundamental principle in the quest to mitigate the negative impacts. An important element in that sense is the sharing of knowledge to improve decision-making concerning various aspects of the built environment, its performance, and potential to reduce the negative effects. More specifically, the effective use and sharing of information have been identified to aid better informed design decisions, more accurate treatment of performance variables and therefore better design outcomes with minimal environmental impact (Abanda et al., 2013).

In that relation, the rapid technological evolution experienced over recent decades has had a radical effect on all aspects of society and its functional mechanisms. The latest developments in Information and Communication Technology (ICT) have opened new doors to cross-domain information and knowledge creation, acquisition and sharing. That also applies to the Architecture, Engineering and Construction (AEC) industry, which has been undergoing a continuous redefinition in terms of collaboration, innovation and digitalisation. The emergence and establishment of Building Information Modelling (BIM) (Borrmann et al., 2018; Sacks et al., 2018) as the most effective collaborative practice has caused a paradigm shift in the perception, use and exchange of building information. There is no universally accepted definition of BIM, but an important point of departure is that it incorporates various processes, methods and data structures over the entire building life cycle to facilitate efficient and accurate creation, exchange and processing of all information related to the built environment. The US National Building Information Modelling Standard (NBIMS-US, 2015) defines the acronym BIM as a three-dimensional matter entailing (1) *"a process for generating and leveraging building data to design, construct and operate the building during its life cycle and allowing stakeholders to have access to the same*

information at the same time through interoperability between technology platforms"; (2) a model encompassing the "*digital representation of physical and functional characteristics of a facility and serving as a shared knowledge resource for information about a facility forming a reliable basis for decisions during its life cycle*"; and (3) management of the process by *"utilizing the information in the digital prototype to aid the sharing of information over the entire life cycle of an asset"* (NBIMS-US, 2015). Advanced BIM practice hereby advises the use of a Common Data Environment (CDE) (British Standards Institute, 2013) for managing information from all stakeholders (Fig. 1-1), including such that is not captured directly in BIM models (e.g. point clouds, design brief documentation, etc.) (Petrova et al., 2018). This increased adoption of BIM technologies and workflows is a part of an important metamorphosis in the industry, which aims for both sustainable modelling of buildings and sustainable management of related information.

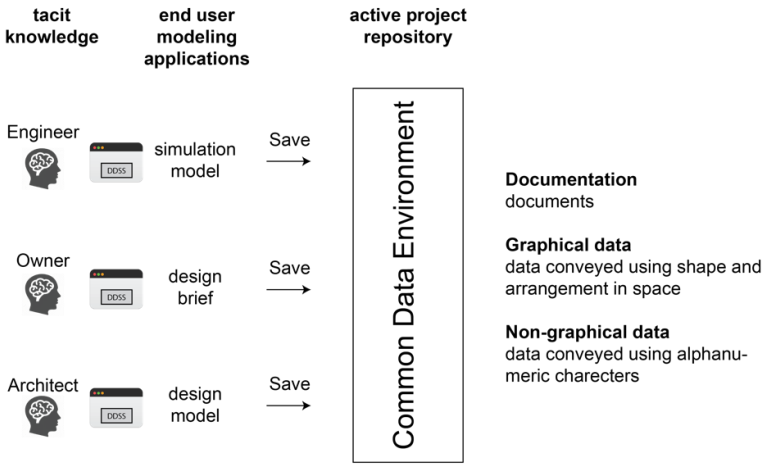

Figure 1-1: Use of a Common Data Environment in collaborative building design (Petrova et al., 2018)

The richness and exponential generation of data during the design, construction and operation of buildings, in combination with advanced technology and analytical approaches have provided the necessary prerequisites for the discovery of valuable hidden insights in the function and behaviour of the existing buildings. Building Monitoring Systems (BMS) and sensor networks hereby allow to track the built environment and provide the valuable input needed to harvest the potential of powerful statistical and symbolic Artificial Intelligence (AI) approaches (Fig. 1-2) (Minsky, 1991; Hoehndorf & Queralt-Rosinach, 2017) such as machine learning, semantic queries, inference machines, etc. (Petrova et al., 2019). Such insights are of significant importance to sustainable design, which aims to incorporate aesthetics and architectural value together with energy efficiency, indoor environmental quality,

occupant comfort, health and productivity into one high-performing whole (Petrova et al. 2018; Petrova et al., 2019).

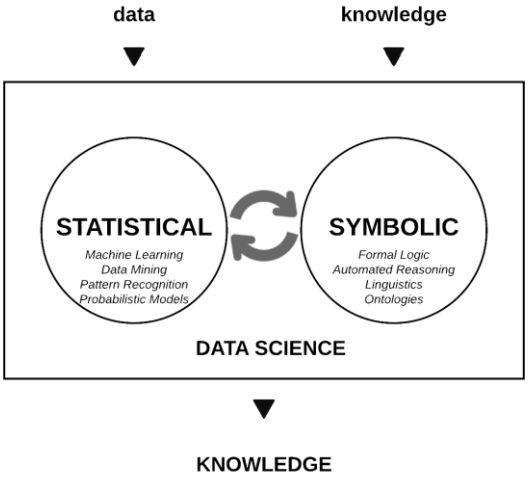

Figure 1-2: Statistical and symbolic constituents of data science and Artificial Intelligence, based on Hoehndorf & Queralt-Rosinach (2017) (Petrova et al., 2019)

More importantly, the combined potential of the different computational approaches and technologies allows to both continuously discover novel knowledge hidden in the operation of existing buildings and document it in a shareable, reusable, modular and extensible way. Such a dynamic knowledge ecosystem spanning across the areas of BIM-based sustainable design and AI can help enhance decision-making and thereby help define, create, monitor and continuously boost the performance of the buildings of the future (Petrova et al. 2019).

However, performance issues remain characteristic to the built environment despite the technological advancement and sophistication of computational design tools, predictive models and simulation mechanisms in support of sustainable design. For instance, reducing the gap between designed and measured building performance has become a central subject in academia, and research indicates that its root causes are attributed to multidimensional reasons spread over the entire building life cycle (de Wilde, 2014). And while some discrepancy is inevitable, research has identified that measured energy use can be as much as 2.5 times higher than the predicted one, which testifies to both its significance and magnitude (Menezes et al., 2012). Even though the AEC industry focuses mostly on the performance gap related to energy use, discrepancies between predicted and actual indoor air quality, thermal comfort, acoustic performance, daylight levels, etc., are also highly likely to occur. That undermines the credibility of the AEC sector and introduces general scepticism towards the concept of high-performance buildings (de Wilde, 2014). During the

design process, such performance discrepancies can be caused by (i) miscommunication concerning performance targets and lack of collaboration between the parties (Carbon Trust, 2012), (ii) inability to accurately predict future use, as well as changes in building operation and occupancy (Menezes, 2012), (iii) inadequate design concepts, assumptions related to analytical input parameters or over-/underestimations (de Wilde, 2014) or (iv) lack of technical detail and buildability of the design solutions (Zero Carbon Hub, 2010).

Considering that modelling and simulation tools are essential predictive and analytical components, their incorrect use is identified as another main contributor to the performance gap (Menezes, 2012; Carbon Trust, 2012). However, it is important to note that their correct use by itself is also insufficient. Domain expertise and capability to choose and apply the right methods the right way, as well as accurate data definitions and input are also required (Dwyer, 2013). Furthermore, according to the Zero Carbon Hub (2010) report, *"Calculations and modelling are often divorced from design and the mechanisms for ensuring that modelling is an accurate reflection of what is built are weak"*. Moreover, BIM models and simulation models are rarely revisited or reused in the operational building stage, and similarly, the design assumptions and concepts remain isolated in the design phase and are never modified based on the actual building performance (Petrova et al., 2019). Finally, the lack of data integration and sharing across domains also plays a significant role in the existence of the performance gap (Hu et al., 2016).

The impact of the issues mentioned above becomes stronger by the use of rules of thumb and previous experiences (Heylighen et al., 2007) as a sole basis in decision-making related to design approaches and parameters, instead of sound evidence (Petrova et al., 2018). Defined as tacit knowledge, such experiences are valuable, but hard to capture and formalise, and are context-specific (Polanyi, 1958; Polanyi, 1966). The increase in experience increases the complexity of tacit knowledge, which evolves into strong design patterns (Alexander, 1977). These patterns are the essence of domain expertise and are highly influential to the design process, yet, alone, they cannot provide the same integrity as an evidence-based system. Their significance, however, can be boosted by external evidence found in the existing buildings. That realisation has also led to in-depth investigations of evidence-based practice for the built environment (Criado-Perez et al., 2019; Hall et al., 2017).

In that relation, symbolic representations and explicit knowledge bases can be used to boost both machine learning approaches and enhance human decision-making, allowing them to create more intelligent sustainable design solutions and build trust to the taken decision. Thus, knowledge discovered in the behaviour of existing buildings and the related design archives can inform future design decision-making, thereby leveraging the multiplicity and richness of the various data sources and paving the way towards evidence-based design practice. Yet, how to close the loop from building operation to design and use that knowledge cycle to provide effective

performance-oriented decision support to the design team has not been explored in detail and is, therefore, the subject of investigation in this thesis.

1.2. ON THE POTENTIAL OF KNOWLEDGE DISCOVERY, REPRESENTATION AND RETRIEVAL FOR SUSTAINABLE DESIGN DECISION SUPPORT

Advanced knowledge discovery approaches allow to discover high-level knowledge in low-level data (Fayyad et al., 1996) and obtain valuable insights in building performance (Fan et al., 2018). Using such knowledge can allow higher level analyses and redefine the way buildings are designed. However, the interpretation, contextualization, reinfusion of the discovered knowledge into future designs and enabling its reuse are fundamental to achieving evidence-based design decision support (Petrova et al., 2019).

In this regard, a reconciliation of statistical and symbolic AI approaches can be of utmost value. Statistical methods are useful in learning patterns or regularities from data, whereas symbolic representations are designed to explicitly capture the knowledge within a given domain and allow various forms of deductive inference (Hoehndorf & Queralt-Rosinach, 2017). Thus, integrating machine learning approaches for knowledge discovery with semantic data modelling and high-level decision support systems in a cohesive context-aware and user-centred ecosystem of rich knowledge bases can be a significant step towards a refined sustainable building design process. Knowledge Discovery in Databases (KDD) and data mining (Fayyad et al., 1996) allow the discovery of novel insights from the large datasets generated throughout the entire building life cycle. Semantic web technologies and linked data allow to formally represent the built environment and retrieve knowledge according to domain-specific requirements (El-Diraby, 2013; Pauwels et al., 2017). Due to their ability to support decision-making, both approaches have independently received major attention in AEC. Combining both can enrich data mining processes with domain knowledge (Ristoski & Paulheim, 2016) and facilitate knowledge discovery, representation and reuse (Petrova et al., 2019a).

Semantic (knowledge) graphs and their ability to represent relations (Sowa, 1992) between buildings, locations, spaces, and other heterogeneous data can scale and articulate the discovered knowledge of how the existing building stock performs in a machine-readable form. Thus, they provide the necessary infrastructure for knowledge reuse and decision support. Graphs can support human decision-making in various ways. The semantic links between the data allow to disambiguate and add context, which is the essence of any knowledge-based system (Sowa, 1992; Sowa, 2008). In other words, semantic graphs allow contextualization of disparate building data and machine-readable articulation of the rich semantic links between them. Therefore, to work towards building performance knowledge contextualisation and demonstrate the value of semantics, collected building data needs to be treated not only in depth with elaborate statistical models and data mining algorithms but also in

breadth to capture the evolution of the discovered knowledge over time. That includes the relation of the building performance insights to other relevant AEC knowledge and the rest of the world context. Thus, blending symbolic and statistical approaches can help achieve holistic and multi-faceted design decision support, which cannot be achieved with any of the approaches independently.

Of course, data availability, diversity, volume and richness are essential to the discovery of actionable insights useful for performance-oriented decision support. Yet, the rapid increase in the volume of data does not automatically guarantee new insights and advances in the understanding of the data (Lausch et al., 2015). To uncover those insights, a continuous flow of the right data to the right analytical mechanisms and actors, in the right format is needed. So far, research has mostly focused on the methods for analysing raw data. Equally important, however, should be discovering how to break up the isolated data silos, how to retrieve data effectively to allow user-centred decision support, how to enable the exploration of unfamiliar datasets from different domains, how to meaningfully reuse and integrate heterogeneous datasets, how to understand and disambiguate data, and how to make data readable and understandable by machines and humans (Janowicz et al., 2015). In other words, turning data into valuable, actionable insights requires an infrastructure not only for analysing data with statistical approaches, but also for publishing, storing, retrieving, reusing, and integrating data, which semantic approaches excel at.

1.3. OBJECTIVES OF THE THESIS

In summary, closing the loop between building operation and design to provide evidence-based decision support in a BIM-based sustainable design process requires (1) understanding the different building data types and (2) appropriate data analysis approaches, as well as defining a (3) clear knowledge discovery goal to be able to further understand the (4) KDD output and representation needs. Furthermore, enhancing the end user's ability to take decisions in the right context also requires (5) interpretation of the discovered knowledge by the use of domain expertise and having (6) a solid user-centred mechanism that allows its access and reuse in a BIM environment.

While KDD and semantics individually may not be sufficient to close the desired cycle, the thesis aims to demonstrate that extending and integrating them makes it possible to discover valuable hidden knowledge in the operation of the existing building stock and unlock its reusability and decision support potential. Therefore, the main research question that this thesis aims to answer is defined as follows:

How can knowledge discovery, representation and retrieval be fused to establish a feedback loop from building operation to design and inform sustainable BIM-based design decision-making in an evidence-based and user-centred way?

To utilise building performance effectively as both hidden knowledge source and decision-making informant, all parts of the design-operation-design cycle need to be investigated. The targeted evidence-based system can only be possible if all parts are effectively functioning and dynamically linked together. Therefore, to materialise the holistic approach and close the loop between building operation and design, the thesis aims to fulfil the following objectives:

(1) Provide a framework for performance-oriented design decision support relying on BIM, data mining and semantic data modelling, thereby allowing customized information retrieval according to defined design goals.
(2) Demonstrate how a semantic cloud of building data enriched with performance patterns can be used by design teams as a knowledge base in decision support.
(3) Showcase how the knowledge can be brought back to design professionals through the design aids they use empowered by user-centred context-aware recommendations relying on an ecosystem of rich knowledge bases.

Fusing the different areas of AI for design decision support can enhance human decision-making and help understand the metabolism of buildings and their occupants. Most importantly, using knowledge discovered in the existing building stock can revolutionize the way we design the buildings and can transform building design from human-centred to humanity-centred.

1.4. THESIS OUTLINE

Chapter 1 outlines the **main background and challenges** in the research domain, as well as the main **objectives** of the PhD research project.

Chapter 2 presents a **state of the art review** in the areas of knowledge discovery in databases, semantic data modelling and user-centred design decision support systems relying on knowledge bases, from both general and building performance improvement perspectives.

Chapter 3 details the various sources and types of building data and demonstrates the methods used for **knowledge discovery, representation and retrieval** for building design decision support.

Chapter 4 investigates the use of **crowdsourcing techniques for interpretation of knowledge** discovered in operational building data and embedding of domain expertise in the knowledge base for design decision support.

Chapter 5 describes the proposed **context-aware design decision support system**, which uses the rich knowledge bases to provide recommendations to the end-user.

Chapter 6 highlights the main **conclusions** of the research.

Chapter 7 provides recommendations for **future work** on the topics investigated in this PhD project.

Appendices A-F contain the collection of the published **journal and conference articles**, as well as the journal articles currently under review that refer to the results from this PhD project.

CHAPTER 2. STATE OF THE ART

"Scientific knowledge is in perpetual evolution;
it finds itself changed from one day to the next"

Jean Piaget

Research in the area of high-performance building design shows that some of the most fundamental success criteria are related to the consistent and dynamic integration of building performance predictions, modelling and simulations in a research-based and data-driven process (Aksamija, 2012). Goldman & Zarzycki (2014) state that BIM can facilitate knowledge transfer between projects, but achieving a holistic standpoint requires to reuse experience from previous projects. In other words, building operation needs to inform the design phase, which can significantly refine the outcome and make it possible to improve decision-making with quantified knowledge in a structured way (Petrova et al., 2018). According to Isikdag (2015), such a future transformation would require enabling an integrated environment of distributed information, which is always up to date and open for the derivation of new information. Furthermore, Goldman & Zarzycki (2014) state that future information exchange in building design also has to be based on reuse of experience across designers, which requires knowledge to be modular and shareable.

As stated in the introductory chapter, the purpose of this study is to pave the way towards such transformation with performance-oriented design decision support relying on BIM, data mining and semantic data modelling for the creation of rich knowledge bases allowing customised user recommendations. Therefore, the following chapter presents a state of the art review of the fundamental building blocks of the research, namely KDD, semantic data modelling and knowledge-based design decision support.

2.1. KNOWLEDGE DISCOVERY IN DATABASES

From an analytical perspective, in-depth research has been performed to identify how to transform data into insights and thereby eliminate drowning in the multiplicity of generated, but unused (building) data. That includes identification of the different types of data, as well as the methods for data selection, preparation and mining. Much of this research employs various machine learning approaches for KDD, which was defined as the overall iterative process of extracting useful knowledge from data by Fayyad et al. (1996). That definition builds on the concept of knowledge as an end product of a data-driven discovery (Piatetsky-Shapiro, 1991) and assumes five main steps, which aim to transform raw data into actionable knowledge of immediate value to the end user, i.e. data selection, pre-processing, transformation, mining and evaluation of the results (Fayyad et al., 1996) (Fig. 2-1):

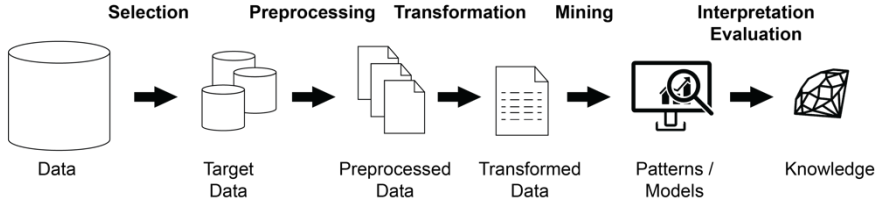

Figure 2-1: The process of Knowledge Discovery in Databases as defined by Fayyad et al. (1996) (Petrova et al., 2018).

As a fundamental part of that process, data mining is defined as the step that employs specific algorithms to discover useful and previously unknown patterns in the data (Fayyad et al., 1996). Hand et al. (2001) later extend that definition to *"the analysis of large observational datasets to find unsuspected relationships and summarize the data in novel ways so that data owners can fully understand and make use of the data."* Seminal in this regard is the work by Bishop (2006) who states that *"pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories".* A more recent study describes data mining as a process of discovering knowledge and patterns in large amounts of data, thereby indicating that the data sources can include databases, warehouses, the Web, other repositories, or data that are dynamically streamed into the system (Han et al., 2012). The authors further extend the KDD process by adding two steps, namely data integration, which allows combining multiple datasets, and use of visualization and knowledge representation techniques to present the mined knowledge to end users. The most interesting patterns may then be stored as new knowledge in a knowledge base. Patterns are classified as interesting if they are easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validate a hypothesis that the user wanted to confirm (Han et al., 2012).

Even though data mining is the step that directly contributes to the identification of valuable patterns in the data, research underlines the significant importance of data preparation to the KDD process (Soibelman & Kim, 2002). Cabena et al. (1998) indicate that 60% of the time is attributed to data preparation, whereas data mining itself accounts for only 10% of the total effort.

2.1.1. KNOWLEDGE DISCOVERY ACCORDING TO PURPOSE AND DATA TYPE

Based on the purpose of knowledge discovery, Fayyad et al. (1996) define six widely accepted data mining categories, namely classification, clustering, Association Rule Mining (ARM), regression, summarization and anomaly detection. Han et al. (2012) later extend that definition and outline the following main data mining functionalities: characterization and discrimination; mining of frequent patterns, associations and correlations; classification and regression; clustering analysis; and outlier analysis.

Each method/functionality belongs to one of two main categories, i.e. predictive and descriptive. Predictive techniques rely on sets of observations with given input and output variables (training data) and use statistical models and forecasting approaches to predict the future and provide actionable insights. The predefined inputs and outputs, however, make the discovery of novel knowledge unlikely (Han et al., 2012). A classic example here is image classification, which relies on annotated images used as training data to be able to classify previously unseen images according to the given class labels.

Descriptive analytics, on the other hand, are quite powerful when it comes to discovery of the intrinsic structure, correlations and associations in data and can uncover previously unknown and hidden knowledge (Han et al., 2012). In other words, while predictive analytics adopt a backward approach by having a predefined target, descriptive methods are forward oriented and help discover interesting relationships that bring out the value in the data (Fan et al., 2018). When the objective is to discover novel knowledge and the laws governing the relationships between building data parameters (e.g. regularities or irregularities in sensor data), it is necessary to use descriptive approaches. That is because any data selection, predefined input/output, parameter determination or limiting the number of input variables and expected results (characteristic features of predictive methods) limits the possibilities of identifying performance insights, correlations and novel discoveries related to the intrinsic structure of the data (Fan et al., 2018).

In terms of the input data source, research defines several categories. Han et al. (2012) distinguish mainly between database data and transactional data but elaborate further that data mining techniques can also be applied to data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and web data. Similarly, Lausch et al. (2015) also outline numerical and categorical data, text, web, media, time series and spatial data as main categories in terms of input data. In AEC, there is a prevalence of building data with spatio-temporal character, e.g. data linking building objects in a given location (BIM models) to recorded observations at a given time (time series data from sensor networks) (Petrova et al., 2019a). Time series data is usually defined as a collection of chronological observations, which are large in size, high in dimensionality and updated continuously (Fu, 2011). Fu (2011) also states that knowledge discovery in time series usually targets the extraction of events, clusters, itemsets, motifs (frequent sequential patterns), discords (infrequent sequential patterns), anomalies and association rules. Spatio-temporal data are central to the context of this research effort as they could capture both physical properties of the buildings, design rationale and real-time performance, thereby nurturing the holistic approach to decision support.

2.1.2. KNOWLEDGE DISCOVERY FOR BUILDING PERFORMANCE IMPROVEMENT AND DESIGN DECISION SUPPORT

Both predictive and descriptive methods have received major attention provoked by the need for improving building performance, enhancing sustainability and bridging the performance gap, and the availability data (Yu et al. 2016; Molina-Solana et al., 2017; Bilal et al., 2016). That has resulted in a significant body of literature examining the vigorous use of knowledge discovery for decision-making and building performance improvement.

An analysis of the scientific literature landscape (Petrova et al., 2019; Petrova et al., 2018) indicates that research employs **predictive** data mining approaches mainly for forecasting of energy demand and aiding energy savings (Ahmed et al., 2011; Zhao & Magoules, 2012; Wang & Srinivasan, 2016; Amasyali & El-Gohary, 2018; Ahmad et al., 2018; Li et al., 2018), prediction of building occupancy and occupant behaviour modelling (Zhang et al., 2018; D'Oca et al., 2018; Chen & Soh, 2017), as well as fault detection and diagnosis of anomalous behaviour in building systems (Cheng et al. 2016; Kim & Katipamula, 2018).

The use of **descriptive** methods in research is usually associated with building energy management (Fan et al., 2018), framework development (D'Oca & Hong, 2015; Fan et al., 2015, Yu et al., 2013), understanding occupant behaviour (D'Oca et al., 2018), improvement of building operation (Xiao & Fan, 2014), and extraction and understanding of patterns in energy use (Miller et al., 2015).

Other highly relevant categories that combine different methods include model calibration and improvement of design and simulation input (Kim et al., 2011), and design pattern extraction (Yarmohammadi et al., 2017; Tucker & de Souza, 2016). State of the art related to each of those aspects and the interrelations between them are discussed in the following sections.

Anomaly detection and building diagnostics

Several researchers efforts have highlighted as fundamental the importance of understanding the behaviour of buildings to be able to predict anomalies and faults in building operation and thereby improve performance (Fan et al., 2018; Pena et al., 2016; Fong et al., 2018; Zhu et al., 2018). Capozzoli et al. (2018) claim that faults in building operation (e.g. HVAC systems, equipment, building control systems, etc.) significantly contribute to the performance gap. Therefore, the authors state that preventative data-driven measures such as characterizing energy consumption patterns over time are of high importance (Capozzoli et al., 2018). In that relation, Fan et al. (2015) demonstrate the potential of temporal knowledge discovery in operational building data by using energy consumption pattern clustering and ARM for detecting anomalous system operation, preventing deficit flow and thereby improving building performance.

Design and energy performance optimization

In terms of improving decision-making, research often points to the use of knowledge discovery methods for enhancement of energy efficiency as a determinative attribute of building performance and sustainability. For instance, Fan et al. (2018a) rely on gradual pattern mining for determination of co-variations between numerical building variables with high influence on performance. Furthermore, in a recent effort, Fan et al. (2019) propose a framework, which uses interpretable machine learning approaches to assist in explaining and evaluating energy performance models and ultimately eliminating inaccurate predictions. In general, the use and benefits of descriptive analytical techniques for knowledge discovery in operational building data have been discussed at length (Fan et al., 2018; Miller et al., 2018). Miller et al. (2015) address automation as another essential perspective related to the efficiency of data mining tasks for extraction of building performance insights from large unstructured datasets. Miller et al. (2018) further consider a crucial part of that process, namely supporting the human interpretation of the data mining results with visual analytics. Cebrat & Novak (2018) use clustering methods to elaborate on the relationships between building energy parameters affecting energy performance, thereby expanding the knowledge related to the choice of optimal design parameters. In that relation, Zhang et al. (2018) address the correlations between building features (building physics, weather conditions, occupant behaviour) and data mining output, and the impact of feature engineering on the accuracy of machine learning algorithms for building energy data mining. In another effort focusing on decision support, Geyer et al. (2017) use clustering methods to determine building retrofit strategies while focusing on cost-efficiency.

In sustainable design practice, several efforts rely on data mining for decision support in the definition of sustainability certification objectives (Jun, 2017; Kim, 2017). In terms of enhancing predictions, Ahmad et al. (2017) compare different models for forecasting of energy demand to determine variations in accuracy and efficiency. Son & Kim (2015) use data mining techniques and early-stage project variables to predict the performance of green buildings. When it comes to predictive decision support mechanisms, research shows successful implementation of knowledge discovery approaches for classification of factors influencing primary energy demand and evaluation of design variables, which need to be considered during the design process (Capozzoli et al., 2015). In that sense, Mason & Grijalva (2019) demonstrate the potential and latest advancements in sensor technologies, advanced control algorithms and reinforcement learning for the development of autonomous building energy management systems and enhancing building performance. In another effort, Capozzoli et al. (2017a) propose a data mining methodology for defining decision-making rules to identify energy consumption patterns in residential flats and evaluate potential retrofit results. Ashouri et al. (2018) aim for keeping the human in the loop and investigate the use of data mining for analysis of historical energy use data and reducing energy consumption by recommendations to the building occupants.

Building occupancy and occupant behaviour

With regards to building occupancy, research underlines that occupant behaviour is critical to building performance due to its highly unpredictable nature. Crucial in that relation are behavioural patterns related to window opening, lighting control and space heating/cooling (Sun et al., 2019). D'Oca et al. (2018) stipulate that understanding such behaviour is critical for enhancing energy performance, reducing operating costs, improving indoor environmental quality and occupant comfort, etc. To harvest those gains, a significant research effort is dedicated to deciphering occupant behaviour, including identifying the most appropriate methods for data collection and most accurate behaviour modelling techniques (Sun et al., 2019; D'Oca & Hong, 2015; Capozzoli et al., 2017; Wolf et al., 2019).

Model calibration

As a core concept in this study, the reuse of measured performance data for informing and improving building design and the associated decision-making processes have, to some extent, been addressed in research. The performed extensive literature review identified that such efforts usually relate to the use of measured performance data to improve the accuracy of design input, simulations and thereby output. For example, Garrett & New (2015) utilise measured energy consumption data for autonomous tuning of building energy models. Tronchin et al. (2018) also use monitored building data for continuous model calibration together with parametric simulation to increase the robustness of performance estimates in the design phase. Several researchers also address calibration of building energy models to measured data through data mining and/or evidence-based methodologies (Lam et al., 2014; Mihai & Zmeureanu, 2013; Raftery et al., 2011).

Design pattern mining

When it comes to knowledge discovery and reuse associated with data originating in the design phase, research usually targets pattern discovery in BIM and simulation data for decision-making support. For instance, Jin et al. (2018) use clustering and feature extraction approaches to retrieve spaces with similar usage functions. The authors compose a method for automatic learning of spatial design knowledge from Industry Foundation Classes (IFC) data based on boundary graphs with space boundary relationships (Jin et al., 2018). Liu et al. (2015) develop a data-driven workflow for energy efficient building design to improve the accuracy of performance analyses and reduce the time for completion of design iterations. They aim to integrate a logical workflow informed by data mining results in the integrated design process and discover the best correlation between different energy systems in BIM models. Thus, Liu et al. (2015) clearly respond to the idea of using knowledge discovery methods to support decision-making in a performance-oriented design process. Pattern discovery in design data is further discussed by Yarmohammadi et al. (2017) who aim for extraction of 3D modelling patterns from BIM log text data. Peng et al.

(2017) use insights discovered in BIM data to provide recommendations for improved efficiency in the maintenance phase and better resource use.

Few researchers address the reuse of knowledge discovered in data for design decision support from a more holistic perspective. Tucker & de Souza (2016) actively investigate the use of building performance simulation patterns for the creation of a "repository of knowledge" for design decision support. Their research builds on earlier results presenting a framework for design decision-making, which adopts a user-centred approach and considers the sequences of design actions that novice designers undertake (de Souza & Tucker, 2015). Furthermore, de Souza & Tucker (2016) propose a conceptual data model, aiming to present dynamic thermal performance simulation information to designers in a meaningful way, thereby supporting decision-making. In another effort aiming to highlight the potential for knowledge discovery in BIM data, Krijnen & Tamke (2015) investigate the potential of machine learning approaches for the extraction of implicit knowledge from BIM models and the possibilities that such an approach can provide for machine-readable qualitative description of buildings.

Drivers for data mining applications in the AEC industry

And while the state of the art review testifies to a diverse potential of KDD approaches for design decision support and building performance optimisation, that potential itself has become the subject of various studies, aiming to assess the actual usefulness of KDD to the AEC industry (Ahmed et al., 2018; Gajzler, 2016; Gajzler, 2010). For instance, Ahmed (2018) investigate the current challenges and drivers for the use of data mining approaches in the industry and report that sustainability and decision support systems are among the six main drivers. They summarize that strongest potential for data mining is found within design, construction, sustainability and energy analysis, forensic analysis and reuse of digital building components. The authors' findings state that when it comes to the design process, data mining applications are recognised as potentially most useful for creating a feedback loop from building operation to design (Ahmed et al., 2018).

Even though the performed literature study has identified a certain level of recognition of the powerful potential of KDD approaches in sustainable design practices, the AEC domain is still lacking some fundamental advancements that would deploy its full potential for design decision support. A common element in all studies, regardless of the adopted methods, algorithms and their level of sophistication, is the need of human expert interpretation of the results and appropriate infrastructure that would allow the reuse of the discovered knowledge. The need for a feedback loop between building operation and design is also recognised; however, bridging those phases in a holistic and circular manner has not been explored in detail. Even though research recognises the potential benefits of evidence-based decision-making and the necessary prerequisites have been discussed, holistic approaches have not been successfully implemented. As stated by Petrova et al. (2019) and as seen in the state of the art

review, the existing solutions usually make use of KDD to support decision-making processes in the same phase where the data originates from. In other words, improving decision-making in the design phase usually relies on insights discovered in BIM models and simulation data, while mining measured performance data often aims to improve building energy management and operational performance. Research shows that knowledge reuse across the phases of the building life cycle is usually related to building energy model calibration for improvement of the simulation output, which, despite responding to the general idea, only partially fulfils the target of establishing a feedback loop between design and operation.

2.1.3. CHALLENGES AND LIMITATIONS

Even though KDD approaches allow the discovery of valuable building performance insights, several important limitations need to be addressed. Traditionally, data mining techniques are usually applied on data batches, i.e. isolated silo data. Unless data mining happens in an automated real-time fashion, optionally including stream processing technologies (Della Valle et al., 2009), the discovered insights and the related conclusions remain limited and do not address complexity, interdisciplinarity and continuous generation of data (Lausch et al., 2015). Also, data selection, pre-processing, cleansing and variable selection resides in the hands of the human analyst, who is responsible for the fitting of the data to the knowledge discovery goals and the needs of the data mining algorithms. The subjectivity related to the human factor is hereby directly influential to the results, which in cases of inaccurate decisions may become the reason for false positives or neglecting novel knowledge (Petrova et al., 2019).

Han et al. (2012) also outline the major issues related to data mining research and partition them in five main categories related to mining methodology, user interaction, efficiency and scalability, diversity of data types, and societal aspects. The authors state that besides considerations related to mining in multidimensional spaces, use of multidisciplinary approaches and semantic relationships, data mining approaches should also more strongly consider noise, uncertainty and incompleteness of data to ensure accurate results. Another major issue, also strongly highlighted in this thesis, is the incorporation of the end user's knowledge in the mining, as well as visualisation, interpretation, and comprehension of the results. According to Han et al. (2012), these aspects are particularly crucial, especially in interactive KDD processes. Of course, the efficiency and scalability of the data mining methodologies and algorithms, as well as handling of complex data types and dynamic repositories are significant issues that are highly relevant in a world of data with exponentially growing versatility, volume and velocity. Finally, a set of ethical considerations and potential impacts in terms of privacy and potential misuse also need to be continuously addressed (Han et al., 2012).

2.2. SEMANTIC DATA MODELLING

Being able to reuse KDD outputs from building operation and inform future design decision-making in a holistic way requires a robust context-aware infrastructure allowing to automate the residence, interpretation and reuse of the discovered knowledge in a dynamic cycle (Petrova et al., 2019). Therefore, the following section addresses the developments in the symbolic representation of knowledge and formalisation of meaning, i.e. semantics. Most of these developments have been in the context of the World Wide Web, which, besides focusing on multi-faceted information exchanges and retrieval, has also been a driver for evolutions related to the semantic representation of objects/datasets and the relationships between them (Bizer et al., 2009).

2.2.1. SEMANTIC GRAPHS, LINKED DATA AND THE WEB OF DATA

In their seminal work, Bizer et al. (2009) discuss the transformation of the World Wide Web into a Web of Data (Linked Open Data cloud[1]). That transformation is expressed by lowering the barriers to publishing, sharing and accessing information on the Web by the use of Linked Data best practices[2]. The term Linked Data was introduced by Berners-Lee (2006), who outlined a set of rules for publishing data on the Web so it becomes an integral part of a single global data space (Bizer et al., 2009).

These so-called Linked Data principles are defined as follows: " (1) *Use URIs as names for things; (2) Use HTTP URIs so that people can look up those names; (3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL); (4) Include links to other URIs, so that they can discover more things.*" (Berners-Lee, 2006). These best practices form the basis of the 5-star open data[3], which assumes defining data according to the Resource Description Framework (RDF) (Grant & Beckett, 2004, Manola & Miller, 2004) data model and linking it with other available RDF datasets, thereby contributing to the LOD cloud.

The RDF data model[4] encodes data in a subject, predicate, object triples (Fig. 2-2). The subject and object constitute the nodes of a graph and can be Internationalized Resource Identifiers (IRIs) (a new protocol element, an upgraded version of the URIs based on Unicode), string literals or blank nodes. IRIs and string literals identify resources ("something in the world"), whereas blank nodes are typically used as mechanisms in defining relations, without representing a specific concept. The predicate of the triple is also represented by an IRI and specifies how the subject and

[1] http://lod-cloud.net/state/

[2] https://www.w3.org/TR/dwbp/

[3] http://5stardata.info/

[4] https://www.w3.org/TR/rdf11-concepts/

object are related (Bizer et al., 2009). The Web of Data is, therefore, a composition of directed labelled graphs. These semantic graphs rely on the triple structure and target uniform syntactic and semantic description of information, making it reusable by both humans and machines.

The Web of Data relies on ontologies (vocabularies), defined as *"formal, explicit specifications of shared conceptualizations"* (Gruber, 1993). They are collections of classes and properties that can be used to describe entities and how they are related (Bizer et al., 2009). Ontologies are also expressed in RDF, using terms from the RDF Schema (RDFS) (Brickley & Guha, 2004) and Web Ontology Language (OWL) (McGuinness & van Harmelen, 2004), which provide varying degrees of expression in modelling. As such, ontologies give meaning (semantics) to the data, with a grounding in Description Logic (DL) (Baader and Nutt, 2003), and allow it to be queried by the use of query and rule languages such as SPARQL and Semantic Web Rule Language (SWRL) (Horrocks et al., 2004). Together, all these fundamental concepts constitute the basic building blocks of the Semantic Web conceived by Berners-Lee (2001) and defined as "*an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."*

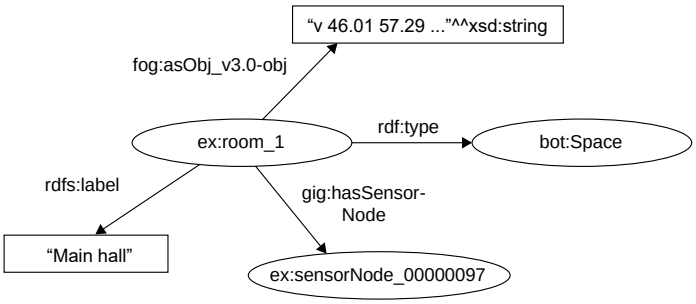

Figure 2-2: A Subject-Predicate-Object triple structure as represented by the RDF data model with ovals representing the subject and object nodes, the arrows representing the predicates and the rectangles representing the literals (Petrova et al., 2019; Petrova et al., 2019a)

2.2.2. LINKED BUILDING DATA

Throughout the last decade, the AEC domain has also recognized the potential of semantic web and linked data technologies. Pauwels et al. (2017) performed an extensive review outlining the development and application progress of semantic web technologies in the AEC industry. In an earlier effort, Abanda et al. (2013) also explored the trends in the application of semantic web technologies in the built environment, including such related to sustainability and energy efficiency. One of the most notable efforts in the area is the early work in transforming IFC into an OWL ontology (ifcOWL) (Beetz et al., 2005; Pauwels & Terkaj, 2016). That initiative laid

the foundation for the creation of the buildingSMART Linked Data Working Group (LDWG)[5] and the W3C Linked Building Data Community Group (W3C LBD CG)[6], which aim for standardization of the representation and exchange of building data over the web.

The ifcOWL ontology is defined according to three main criteria, among which to *"match the original EXPRESS schema as closely as possible"*(Pauwels & Terkaj, 2016). That, however, has resulted in a large ontology that mirrors the IFC schema almost entirely, thus making it complex, difficult to extend and non-modular (Petrova et al., 2019; Petrova et al., 2019a). Therefore, several other efforts (Fig. 2-3) focus extensively on modularity and extensibility and aim to define an ecosystem of smaller, modular and extensible Linked Building Data (LBD) ontologies (Schneider et al. 2018). At the core of this concept is the Building Topology Ontology (BOT) (Rasmussen et al., 2017), which defines and aims to standardize terms as 'Building', 'Site', 'Space', 'Element', etc. and to which alignment from other ontologies can be made (Schneider, 2017). That includes various domain ontologies such as SAREF[7], DogOnt (Bonino & Corno, 2008), PRODUCT (Costa & Madrazo, 2014), Ontology for Property Management (OPM) (Rasmussen et al., 2018), etc.

Another direction in ontology engineering in the AEC domain is related to the representation of 3D geometric data. Such data presents higher level challenges when it comes to representation with linked data techniques and constitutes a separate LBD module. Related research efforts aim at both representation and linking to other types of building and geospatial data (McGlinn et al., 2019).

[5] https://technical.buildingsmart.org/community/linked-data-working-group/

[6] https://www.w3.org/community/lbd/

[7] http://ontology.tno.nl/saref/

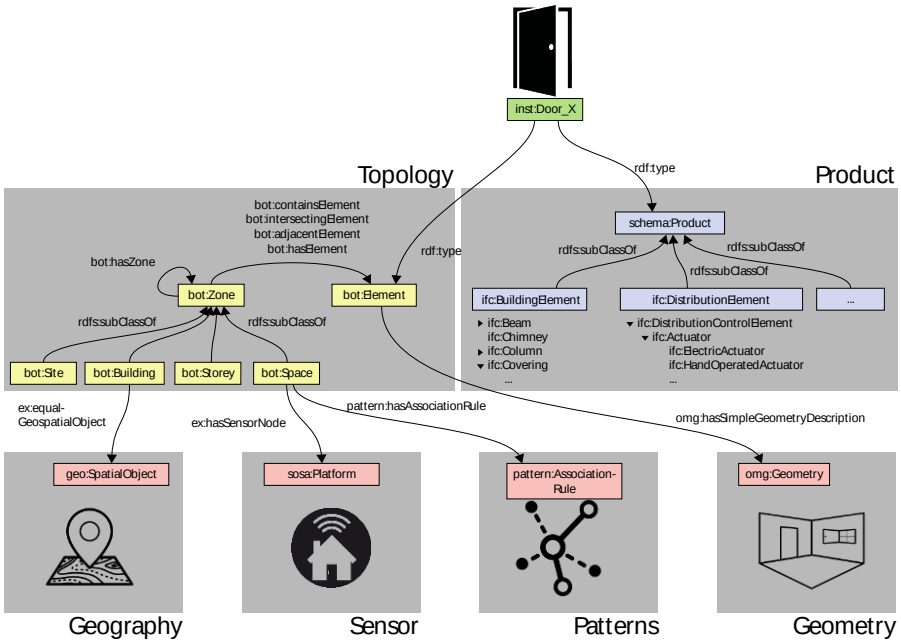

Figure 2-3: Conceptual overview of the modules and ontologies in the linked building data cloud, based on initiatives in the W3C Linked Building Data Community Group and user contributions (Petrova et al., 2019; Petrova et al., 2019a)

2.2.3. SEMANTIC SENSOR DATA

An important body of work belonging to the semantic web and linked data domain and of high relevance to this research resides in the context of sensors and actuators. That is valid from both an analytical and knowledge discovery perspective, as well as from a representation perspective as part of the LBD realm. It is expected that 40% of all data generated globally by 2020 will be from sensors and sensor networks. Thus, the analysis, storage and representation of sensor data have been in the spotlight of research in the past decade. As stated by Petrova et al. (2019a), in the context of the built environment, sensor nodes are placed in precisely determined locations with a dedicated and predetermined purpose of observation, which aims at monitoring building use and performance (occupancy, indoor environmental quality, electricity consumption, etc.) in a real-time manner. This usually results in large amounts of continuous real-time data streams, which are typically captured in optimized for the purpose databases (data lakes) and can serve various purposes related to extraction of behavioural insights from buildings. As such, sensor data constitutes a separate module complementing the LBD cloud (Petrova et al., 2019a).

By the use of dedicated domain ontologies, sensor data can also be stored in RDF graphs, which has resulted in concepts such as Semantic Sensor Networks (SSN) and

Semantic Sensor Web. Ontologies that can be used for this purpose include SAREF, SEAS (Lefrancois et al., 2017), SOSA[8], SSN[9]. In that relation, several research efforts focus on the semantic representation of sensor data to support various aspects in decision-making in the AEC domain (Rasmussen et al., 2018; Petrova et al., 2019; Petrova et al. 2018a; Schneider et al., 2018). The main difference between the approaches can be found in the stance that the researchers take on storage mechanisms for sensor data. For instance, Rasmussen et al. (2018) and Schneider et al. (2018) store collected historical data directly in the RDF graph. Petrova et al. (2018a) take a different approach by maintaining the sensor data in its native storage and embedding a direct link to that location in the semantic graph of the associated building. By following the links from the RDF graph to the web Application Programming Interface (API), sensor data can then be retrieved by the end-user application through on-demand HTTP requests. Such an approach can be valuable in cases where data is continuously collected, and its exploration is based on real-time analytics. Furthermore, due to the high volume and velocity of sensor data, storing large datasets in the RDF graph usually leads to a "swollen" graph (Petrova et al., 2019), which has a negative effect on query and reasoning performance. Therefore, by keeping the raw sensor data outside the RDF graph, in its native storage and format, end-user applications can easily parse the much smaller graph, while still maintaining a live link to the original observations and their numerical values (Petrova et al., 2019; Petrova et al., 2018a).

Another key aspect addressed in research is the heterogeneity of sensor data sources and environments (Calbimonte et al., 2012). Depending on the sensor network and the devices themselves, monitored building data is represented in different ways, with varying data models and underlying schemas. That leads to multiple representation, interoperability and data fusion issues, which have been addressed in several research initiatives. For instance, in an attempt to solve these issues, Sheth et al. (2008) propose to annotate sensor data semantically. Calbimonte et al. (2010) point to the provision of ontology-based access to streaming data as a possible solution and Wang et al. (2015) discuss using SPARQL queries with streaming extensions for direct access to observations. These works aim for reformatting the raw sensor data in a way that allows semantic querying, which requires mapping, annotating and processing data to the alternative semantic representation. Wang et al. (2015) present an extensive overview of semantic sensor net ontologies, mapping and querying mechanisms.

Figure 2-4 summarises the most common means of treating sensor data, including the storage and access mechanisms. Research shows that the collected observations are usually either stored in a database (e.g. an SQL store) or are directly processed using stream processing technologies (Llanes et al., 2016). In both cases, the data is usually

[8] http://www.w3.org/ns/sosa/

[9] https://www.w3.org/TR/2017/CR-vocab-ssn-20170711/

made available for access in a direct API interface. Alternatively, recent research trends suggest processing the raw data to linked data, which, besides providing alternative opportunities for accessing and querying the data, also makes it directly integrable with other semantic data. For instance, Calbimonte et al. (2012) highlight the potential of using SPARQL queries with streaming extensions to access the sensor observations. RDF stream processing and reasoning approaches may provide several different benefits for publishing and analysing real-time sensor data streams and allow to avoid the "swollen" graph issue while at the same time make it possible to include the data in the LBD knowledge graph (Petrova et al., 2019a).

In that relation, Della Valle et al. (2009) state that bringing out the real value of the observations requires a paradigm shift in the consumption of data, i.e. moving from "one-time semantics" (storing data in databases and querying it on demand) to "continuous semantics" (using continuous queries and analyzing the data in a real-time manner). The authors also outline reasoning capabilities over rapidly changing information as a key direction for development. The main transformation stages in the publication of sensor data as RDF streams have been defined by Llanes et al. (2016). They include (1) conversion from sensor data streams to RDF streams, (2) storing the RDF streams, and (3) linking them with other data. These three stages are dependent on the selection of three key elements, namely (i) mapping mechanisms (e.g. D2RQ, R2RML, etc.), (ii) ontologies (e.g. SEAS, SOSA, SSN, etc.) and (iii) the continuous query language with streaming extensions (e.g. C-SPARQL, SPARQLstream, etc.) (Llanes et al., 2016; Calbimonte et al., 2012; Barbieri et al. 2010). An essential aspect is the choice of the additional appropriate datasets to link to so that the potential behind breaking the data out of isolation can be fully harvested.

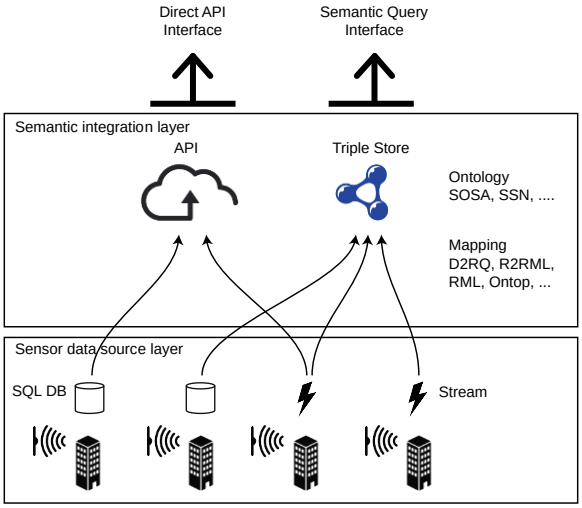

Figure 2-4: An overview of common technical approaches for making sensor data available to an end-user application (inspired by Wang et al.2015) (Petrova et al., 2019)

2.2.4. SEMANTIC APPROACHES TO BUILDING PERFORMANCE IMPROVEMENT AND DESIGN DECISION SUPPORT

As previously indicated, the lack of data integration and sharing is reported as a significant contributor to the performance gap (Hu et al., 2016). Curry et al. (2013) approach this issue with a method for linking of various kinds of building data in a graph of semantic data used for holistic building management. Hu et al. (2016) further highlight a valuable perspective on that integration by stating that linking data that is traditionally kept separate may enable much higher level analyses. For instance, linking occupant behaviour patterns to building operation may redefine the discovery of building performance insights (Hu et al., 2016). Several studies underline that data should be stored in its most appropriate format and linked to create an integrated and well-connected semantic network that can allow the sharing of data in accordance with the end user needs, e.g. keeping sensor data in an SQL store and linking it with contextual semantic data (Petrova et al., 2019; Hu et al. 2016; Curry et al., 2013).

Semantic interoperability between complex systems in building operation and its contribution to building energy performance improvement has also been extensively discussed. Benndorf et al. (2018) confirm that data with a unified structure and meaning should create the basis for interoperability between heterogeneous applications and their associated data representations. Corry et al. (2015) address this issue with a performance assessment ontology and a framework aiming to translate heterogeneous building data into semantically enriched building performance analysis input. To showcase the potential of linked data technologies for minimization of the performance gap, Hu et al. (2018) propose an automated performance evaluation approach relying on integration between OpenMath[10] and linked data to help evaluate performance metrics extracted from sensor data. O'Donnell et al. (2013) target building performance optimization through the combined use of linked data, scenario modelling and complex event processing. Zhong et al. (2018) develop an ontology-based framework for environmental monitoring and compliance checking that integrates building data, environmental sensor data, and regulatory information based on building regulations and design requirements. In terms of semantic unification of data for performance optimization, Diaz et al. (2013) develop an ontology for standard representation of energy efficiency concepts in buildings. The use of semantic web technologies for multi-objective design optimization and energy, environmental and economic building performance has also been investigated (Pont et al., 2015).

Few recent approaches also acknowledge the potential of integrating knowledge discovery and semantic approaches for building performance improvement and design decision support. Esnaola-Gonzalez et al. (2018) present an innovative method for energy efficiency prediction, which combines semantic web technologies and KDD in a smart prediction assistant. In another effort, Esnaola-Gonzalez et al. (2018a) further explore the potential of that combination to ensure thermal comfort in

[10] https://www.openmath.org/

workplaces. In that case, the presented framework is used to support the interpretation phase of the knowledge discovery process, where semantic technologies are used to explain predictive models related to temperature levels as part of thermal comfort regulations that have to be fulfilled (Esnaola-Gonzalez et al., 2018a). McGlinn et al. (2017) also accentuate on the importance of knowledge-based systems and propose an energy management system using Artificial Neural Networks (ANNs), Genetic Algorithms, and Decision Tree rules for building environment optimisation through recommendations alerting to potential energy-saving actions. In a similar approach, Delgoshaei et al. (2018) combine KDD and machine learning techniques to identify energy consumption patterns and store the results in ontologies for further inference. On a more general building level, Szilagyi & Wira (2018) use a similar approach in a smart building context and define a model for a BMS based on hybrid knowledge, which aims to optimize the use of different resources (energy, water, etc.), while still assuring the occupants' comfort. Ploennings & Schumann (2017) showcase the power of integrating computational approaches with an innovative method that uses semantic reasoning to model physical relationships of sensors and systems, machine learning for anomaly detection in energy flow, building occupancy and occupant comfort, and speech-enabled Augmented Reality interfaces for immersive interaction with the networks of devices in the context of a cognitive building. Finally, in a recent effort, Fan et al. (2019a) use graph mining techniques to discover complex relationships in building operation by mining graph data directly.

2.2.5. CHALLENGES AND LIMITATIONS

As seen in the performed review, semantic technologies uncover several higher-level opportunities for holistic design decision-making and building performance optimization. However, some challenges that extend beyond the practical use of linked data and semantic web technologies and are particularly relevant in the current context need to be considered. For instance, as previously mentioned, storing large amounts of sensor observations in the semantic graph may lead to a significantly large ("swollen") graph and affect the overall performance of querying and reasoning, which defeats the purpose of linked data and semantic web technologies. That can be prevented by storing the different kinds of data in their dedicated systems and formats (Petrova et al., 2019).

Another issue outlined in Petrova et al. (2019) that has to be considered is related to the stability of the ontologies used for representation and querying of data. The change of the vocabularies over time means that data has to be reformatted in accordance with these changes, which can result in data loss. Such changes cannot be prevented entirely, yet one should aim to keep ontologies relatively stable, which can be achieved by standardization efforts.

Furthermore, significant challenges arise in terms of the semantic representation and interpretation of the discovered performance insights. For concepts that are semantically explicitly definable, the representation itself may not be an issue, but

once semantics is encoded in the LBD graph, it becomes a static, intermediate, single-perspective view, which needs to be updated in accordance with the dynamic behaviour of the buildings and the continuous data generation.

Since knowledge as a concept occupies a significant space in the objectives of the current research and spans over both the human and machine contexts, these concepts are further discussed.

2.3. KNOWLEDGE-BASED DESIGN DECISION SUPPORT

Significant in the context of this research is the interpretation of the KDD results. As all knowledge takes shape through interpretation, the human input required for interpretation, clarification and disambiguation of machine learning output is vital to the (re)usability of the discovered performance insights. Essential here is the fact that the process of interpretation is to a large extent guided by the background knowledge of the human domain expert, which mainly consists of tacit concepts. The tacit knowledge is accumulated through experience, training, learning and observation, and is deeply embedded in the individual cognitive intuition of the interpreter. That makes this kind of knowledge highly intra-personal and inaccessible unless externalized (Polanyi, 1966; Polanyi, 1958). And even such externalization is only an abstract and limited reflection of the richness of the mind, i.e. an externalized model. In the context of design, the experience and background knowledge are referred to as guiding principles, which are internally embedded and difficult to externalize 'design rules' or 'design patterns', which are deployed by any practitioner, and characterize the way they think (Lawson, 2005).

Hence, the tacit dimension has proven to be virtually impossible to fully formalize and implement in a machine semantically, which presents a challenge when it comes to the embedding of qualitative data and interpretations of building performance in the enriched LBD graph. Yet, expert assessments can, to some extent, be formalized through connection with tangible and explicit performance concepts. That can be invaluable to the disambiguation of performance patterns and can help move closer to semantics- and context-aware decision support. Therefore, to be able to provide context for the further presentation of the contributions of this study, this subchapter considers the aspects and latest developments related to knowledge representation, retrieval, reuse and reasoning for design decision support from both end-user and machine perspectives.

2.3.1. KNOWLEDGE CONTEXTUALISATION AND REASONING- HUMAN VS. MACHINE

The need to utilize data in an effective and meaningful way has made knowledge a focal point in AI, which combines various fields such as machine learning, knowledge representation, ontologies, logic, Natural Language Processing (NLP), reasoning, neurocomputing, etc. As defined by Barr & Feigenbaum (1981), *"Artificial*

Intelligence (AI) is part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit characteristics we associate with intelligence in human behavior- understanding language, learning, reasoning, solving problems, and so on."

In that sense, knowledge and knowledge representation are essential to knowledge-based systems aiming to harvest the potential of the different areas of AI to support human decision-making. The previous section already discussed the tacit concept and its importance to decision-making. Tacit knowledge cannot be expressed directly through vocabularies, but explicit (articulated) knowledge is easy to share and store by means of one or a combination of different ontologies (Sowa, 2008). And while the tacit concepts reside in the human brain and define human knowledge, several fields (cognitive maps, concept maps, semantic networks, ANNs, Hierarchical Temporal Memory (HTM), Knowledge-Based Neurocomputing (KBN), etc.) have emerged that aim to refine knowledge representation and emulate human thinking, problem solving and reasoning. Human reasoning has been assumed to be based on mental logic (reasoning depends on a tacit mental logic, consisting of formal rules of inference) (Beth & Piaget, 1966; Braine & O'Brian, 1998), mental models (creating mental models of the world based on vision, description and personal knowledge and experience) (Johnson-Laird, 1983), abduction (finding the simplest and most likely explanation for observations) (Peirce, 1958), memory-prediction system (remembering sequences of events and their nested relationships and making predictions based on those memories) (Hawkins & Blakeslee, 2004), etc. Thus, understanding human intelligence to be able to achieve machine intelligence has been a subject of investigation of decades of work. So far, that has proven to be impossible, as the concepts of knowledge and reasoning in the context of machines appear to be merely associated with data processing, which does not reflect the higher-level abstractions and processes in the human brain. Pauwels et al. (2012) indicate that one can either aim at implementing autonomous agents mimicking the human reasoning process, which has not been achieved so far, or build engineering applications in assistance of the human decision-maker.

In the context of information systems, Brachman & Levesque (2004) define knowledge as a relationship between a "knower" and a proposition and knowledge representation as *"symbolic encoding of propositions believed (by some agent)"*. Reasoning is, therefore, further defined as *"manipulation of symbols encoding propositions to produce representations of new propositions"* (Brachman & Levesque, 2004). In that sense, KDD methods may be able to provide useful insights in data, and it is possible to manipulate these symbols (data) into new symbols (new data); however, they lack the capability to reason about the meaning and interrelationships between these insights. The reason for that is that KDD relies mostly on statistical models rather than semantic abstractions powered by external knowledge outside these models. In other words, discovered patterns (in building operation) are merely observations, instead of higher level semantic concepts. Contextualising the

patterns could allow to connect them to other externalised semantic knowledge and reason with such knowledge about their meaning, which comes a step closer to human reasoning. Combining the pattern discovery skills of machines with the domain expertise of humans can, therefore, be a valuable resolution to the KDD result interpretation challenge. Machines can be powerful in 'describing' the discovered patterns, which provides valuable input for expert review, interpretation and approval/dismissal. If machines are able to exploit statistics together with semantics and incorporate external knowledge into the reasoning and decision support systems, then a lot of the uncertainty and inaccuracy issues related to design decision-making can be eliminated.

In terms of reasoning in the context of this research, RDFS and OWL concepts enable reasoning to a certain level of complexity (Pauwels et al., 2012). More complex reasoning requires the description of rules with higher-level dedicated rule languages, which enable rule-based reasoning processes (e.g. SWRL (Horrocks et al., 2004), N3Logic (Berners-Lee et al., 2008)). Through the adoption of semantic web technologies, reasoning with data is possible, both in a standard (RDFS, OWL) and more complex (SWRL, SPIN, N3Logic) manner. Instead of going deeper in the topic of reasoning solely, this thesis aims at setting up the overall scene that is needed to be able to deploy advanced reasoning approaches, i.e. combination of data mining and semantics, capturing expert knowledge and building a user-centred system around it that enables its reuse in reasoning.

2.3.2. KNOWLEDGE-BASED SYSTEMS

In support of the research objectives, this section presents an overview of the types of systems that allow the retrieval of discovered and appropriately represented knowledge for design decision support, which is of direct relevance to the main contribution of this thesis. An overall distinction is hereby made between decision support systems (DSS), Case-Based Reasoning (CBR), Expert Systems and Knowledge-Based Systems (KBS), Recommender Systems, and approaches for User-Centered Recommendations.

Decision Support Systems

Existing systems incorporate retrieval approaches that represent knowledge as rules, facts or a hierarchical classification of objects. The related knowledge representation techniques govern the validity and precision of the retrieved knowledge (Malhotra & Nair, 2015). In terms of design decision support, several fundamental concepts need to be considered, e.g. the requirements and (performance) targets, the knowledge (base), the decision support system and the end user.

The processes of acquisition and retrieval of relevant information are essential to any system operating to provide decision-making support. As noted in Petrova et al. (2018), in general, decision support systems (DSS) are defined as computer-based

tools that are adapted to aid and support complex problem solving and decision-making (Arnott & Pervan, 2008; Shim et al., 2002). Power (2002) defines DSS as an interactive computer-based system that assists people in computer communications using data, documents, knowledge, and models to solve problems and make decisions. An important factor discussed in research is the improvement of the efficiency and the effectiveness of the decision-maker (Alter, 2004; Pearson & Shim, 1995). DSS applications entail various sub-technologies and techniques tailored to the decision-making process that they have to support. Haettenschwiler (2001) divides DSS into passive (supports decision-making, but does not provide suggestions/solutions), active (generates suggestions/solutions) and cooperative (allows the decision maker to modify, complete or refine recommendations by the system, before sending them back for validation). Recent efforts also gravitate towards human-centric DSS, which rely on cognitive models to predict human behaviour and adaptive agents to improve system performance (Heytmeyer et al., 2015).

In the AEC industry, research focuses predominantly on the perspectives of ICT and the end user in the development of design decision support systems. Early work discusses the importance and use of DSS for improving communication, knowledge transfer between actors and building life cycle phases and support of a performance-based approach to building performance planning and evaluation (de Groot et al., 1999). More recent works in the performance-oriented design domain aim for the integration of BIM and DSS for sustainable design optimizations, such as the optimal selection of sustainable building components (Jalaei et al., 2015) and evaluation of holistic renovation scenarios (Kamari et al., 2018). Implementations of DSS for facilitating sustainability and buildability assessments (Singhaputtangkul & Low, 2015) and optimal planning of sustainable buildings through the integration of Life Cycle Assessment (LCA) in a DSS (Magrassi et al., 2016) have also been investigated. Chatzikonstantinou & Sariyildiz (2017) propose an alternative decision support framework relying on auto-associative machine learning models that inductively learn relationships between design features of high-performance designs. As the research area pertaining to DSS in AEC is vast and presents a multitude of methods and implementations in accordance with varying decision support objectives in the field, extensive review of implementations is not explored in this work, but the reader may refer to Timmermans (2016) and van Leeuwen & Timmermans (2004) for further details. Many commercial tools (CAD, BIM, simulation, visualization tools, etc.) tools have also been adopted in practice. However, they are mainly standalone applications, which are not built on the principle of knowledge reuse and seldom include the DSS add-ons that offer knowledge from remote resources (Petrova et al., 2018).

Case-Based Reasoning

Reuse of knowledge and experience for decision support, as well as design based on similarity matching has, however, been recognized in research. This is particularly valid for Case-Based Reasoning (CBR), which provides decision makers with a

problem-solving framework relying on recalling and reusing knowledge and experience (Aamodt and Plaza, 1994). Different methods for Case-Based Design (CBD) exist and are differentiated according to their method of implementation. For instance, Dave et al. (1994) implement a system, which aims to aid adaptation and combination of different design cases to support the generation of new designs in a more efficient way. Several research efforts also exploit CBD implementations to support knowledge exchange and renewal between architects (Richter et al., 2007; Heylighen & Neuckermans, 2000). Eilouti (2009) uses design precedents to explore to what extent the reuse of architectural design knowledge is possible.

In terms of high-performance and sustainable building design, Xiao et al. (2017) present an experience mining model, aiming to help the decision-maker find a solution to green building design problems. Shen at al. (2017) combine text mining and CBR to facilitate the retrieval of similar cases in green building design practices. Cheng & Ma (2015) propose a non-linear CBR approach relying on ANN for retrieval of similar certified green building cases. Human-computer collaboration is also explored by Abaza (2008), who targets the creation of a matrix of energy efficient design solutions. To achieve that, the author presents a model, which uses design proposals suggested by a human designer and evaluated by a machine following performance criteria. The combination of CBR and graph matching techniques has also been explored to enable the retrieval of similar architectural floor plans in the early design phase (Sabri et al., 2017). Ayzenshtadt et al. (2016) present a system that combines case-based and rule-based retrieval to enable the search for architectural designs. Weber et al. (2010) propose a solution for a system for retrieval of sketches from a floorplan repository, which utilizes CBR and shape detection technology. However, as stated by Petrova et al. (2018b), these approaches usually rely on classification approaches or topology graphs for capturing of semantics, which are less complex and rich in comparison with BIM and ontologically demarcated data.

Expert Systems and Knowledge-Based Systems

When it comes to Knowledge-Based Systems (KBS), the early examples usually point to expert systems. KBS and expert systems consist of two main components, namely a knowledge base and an inference engine. The main difference between an expert system and a KBS is contained in how and for what the system is used (Malhotra & Nair, 2015). According to Russel & Norvig (2009), expert systems are usually intended to substitute or assist human experts in resolving a complex problem in a more efficient way by reducing complexity. KBS, on the other hand, provide a structured architecture for explicit knowledge representation (Hayes-Roth & Jacobstein, 1994).

Historically, one of the key challenges in both systems has been the validity and consistency of the entire system. All knowledge and rules need to fit in order for the system scope to be consistent, correct, and complete. As a result, a significant effort is needed to make sure that facts and rules are consistent, correct, and complete. This

has proven to be an incredibly hard engineering challenge, which explains much of the downfall of expert systems in the past; even if one puts a tremendous effort in building complete knowledge bases and corresponding rule sets, it will still be only as good as the externalized explicit labels and concepts, which differ significantly from the broader and much more flexible set of tacit concepts that humans utilize. From a technical perspective, Turban & Aronson (2000) hereby highlight that KBS evolved as knowledge became structured, i.e. information being represented with classes and subclasses, and relations between the classes and assertions represented using instances.

Research efforts in the AEC domain target such implementations in various contexts of the performance-oriented design practice. For instance, Nilashi et al. (2015) present a knowledge-based expert system for assessment of the performance of buildings according to green building rating factors. In another effort, Ochoa & Capeluto (2015) propose an expert system using inference to compensate for uncertain or unknown information in the early stages of projects to perform energy and cost performance assessments. In general, knowledge sharing in AEC has been the subject of several investigations exploring, for instance, the use of storytelling as a catalyst for knowledge sharing between projects, architects, companies, etc. (Heylighen et al., 2007), and technology that supports knowledge capture, sharing and reuse (Fruchter et al., 2004).

Recommender systems

Also significant in the context of knowledge reuse is a research area that draws on CBR and KBS approaches and aims to provide decision support to the end user by dedicated recommendations. In general, recommender systems are defined as *"personalized information agents that provide recommendations: suggestions for items likely to be of use to a user."* (Resnick & Varian, 1997). Resnick & Varian (1997) also define the results from a recommender system as recommendations, or in other words, options worthy of the end user's consideration and a result from an information retrieval system interpreted as a match to a user's query.

Research distinguishes between different kinds of recommendation techniques based on the knowledge source. In some cases, that means the knowledge of other users' preferences, while in others it is ontological or inferential knowledge about the domain, specified by a human expert (Resnick & Varian, 1997). Thus, Burke (2007) summarises four main classification techniques: collaborative, content-based, demographic and knowledge-based (Fig. 2-5). Earlier, Brunato & Battiti (2003) also point to context as another important knowledge source. In that relation, knowledge-based recommender systems provide recommendations based on inferences about the end user's needs and preferences. This knowledge may also include explicit functional knowledge about how particular recommendation features meet the user's needs (Burke, 2000). In terms of information retrieval, Musto et al. (2017) divide recommender systems into content-based and graph-based. Content-based systems

hereby provide recommendations based on direct similarity and graph-based systems link user nodes to user-tailored recommendations.

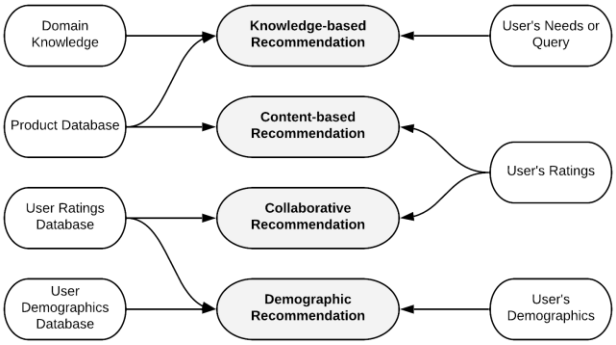

Figure 2-5: Recommendation techniques and their knowledge sources, based on Burke (2007).

In that relation, previous work has investigated recommender systems based on linked data (graph-based) and how they exploit the wealth of data provided by the LOD cloud (Oliveira et al. 2017, Musto et al. 2017). Research on linked data-based recommendations originates in the field of ontology-based recommender systems, first suggested by Middleton et al. (2004). Previous work indicates that the use of linked data and ontologies for disambiguation of content makes recommendation systems semantics-aware (de Gemmis et al. 2015, Boratto et al. 2017). Gorshkov et al. (2016) also use named graphs in a method that allows to ontologically define multiple viewpoints (contexts) in decision support systems.

User-centred recommendations

Passant (2010) was among the first to propose the use of linked data for providing user-centred recommendations. The author makes use of the links between the resources to provide a set of measures for computation of semantic distances between them. Recent works (Oliveira et al. 2017, Boratto et al. 2017) also put the user's profile on focus, thereby usually following an architecture as displayed in Fig. 2-6.

User profiling and suggestion generation is based on user needs and preferences, previous interactions, social relations, likes, etc. The purpose is to match the user's demands (profile) with a recommendation with the highest possible similarity, while still diversifying the recommendations. In the case of building design, for example, if a user indicates high interest in residential nearly zero-energy buildings (NZEB) or LEED Platinum certified buildings, the recommender system should also be able to suggest other NZEB building types, other LEED Platinum building types or other residential building types, etc. (diversification), while keeping the context intact.

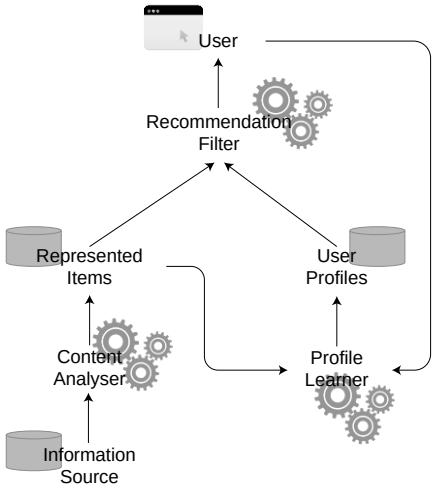

Figure 2-6: Semantics-aware content-based recommender system, based on Boratto et al. (2017) (Petrova et al., 2019a)

Naturally, the richer the dataset, the better the alternative recommendations (Petrova et al., 2019a).

Recommendation systems have, to some extent, been introduced in the AEC industry. However, they are usually content-based systems aiming to suggest predefined objects when a certain level of similarity with the current design is achieved (Petrova et al., 2019a). Research efforts exploiting knowledge graphs and ontologies in a decision support system/recommender system have, however, so far not been pursued in the AEC domain. As part of closing the loop between building design and operation to improve decision-making, one of the fundamental goals of this research is to investigate the application of linked data-based recommendations utilising dynamic building performance knowledge bases in changing context. Critical to that investigation is the understanding of the decisions that have to be taken in a performance-oriented design setting, as well as the relevant data that populates the design-operation loop, its potential contribution and how it can be analysed and reused as new knowledge.

For further details, please refer to Appendix A. Paper I, Appendix C. Paper III and Appendix F. Paper VI: "Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data", "Data mining and semantics for decision support in sustainable BIM-based design" and "Semantic data mining and linked data for a recommender system in the AEC industry".

CHAPTER 3. KNOWLEDGE DISCOVERY, REPRESENTATION AND RETRIEVAL FOR BUILDING DESIGN DECISION SUPPORT

"Man has an intense desire for assured knowledge."

Albert Einstein

As outlined in the previous chapters, this research project aims for design decision support through curated knowledge discovery and representation techniques utilising the variety of data from the built environment and tailored to the needs of performance-oriented building design. The value of the proposed data-driven approach will be highest when it can positively impact both the design decision-making process itself and in turn, the final product (Petrova et al., 2018). Even though the previously discussed AI approaches can empower human decision-making, of high importance is that a data- and technology-driven approach does not neglect the human users and their tacit knowledge, but nurtures it. Bringing out the value of data in the context of this research requires an in-depth understanding of the decisions that have to be taken, the knowledge that they require, the data that has to be collected to discover such knowledge and the suitable knowledge discovery approaches.

Thus, this chapter initiates the effort to close the loop between building operation and design with evidence-based decision support by classifying the most critical decision categories in the sustainable building design process and details the building data types that are directly associated with them. Furthermore, the analytical techniques for each type of building data are outlined. Based on that, the chapter presents the developed research framework and DSS system architecture. Finally, knowledge discovery and representation are performed with collected data from two use cases. The chapter concludes with a demonstration of a user-centred retrieval of the discovered knowledge.

3.1. DATA AND KNOWLEDGE WITH POTENTIAL IMPACT ON DESIGN DECISION-MAKING

Decision-making as a process of finding the best fitting solution to a problem among multiple alternatives to best cater to interpreted objectives has been continuously examined through the years (Simon, 1960; Shim, 2002). That has led to a significant body of work investigating, among others, the phases of the decision-making process,

as well as the models governing it. Such an exploration is beyond the scope of this thesis; however, it is important to achieve an in-depth understanding of both the needs of the end user and the decisions that they need to face in a sustainable design setting. That categorisation is also necessary to be able to provide a definition of the different types of building data that need to be considered and the analytical knowledge discovery techniques that should be used. Thus, as stated by Petrova et al. (2018), an evidence-based approach has the highest impact in scenarios that entail:

- decisions with high impact and criticality, i.e. early-stage design decisions with high level of variability of outcome under high uncertainty;
- specific performance criteria, where the variability of decisions and their practical implications are highly influential to the targeted performance;
- data from a high number of versatile reference buildings;
- data of significant volume, versatility and richness;
- data infrastructure that enables knowledge capture and reuse in decision-making.

3.1.1. SUSTAINABLE DESIGN DECISION-MAKING CRITERIA AND DEPENDECIES

In that relation, an investigation performed by Petrova et al. (2018) indicates that many of the critical decisions related to the early stages of the performance-oriented design process are interdependent and a lot of information concerning their criticality and the criteria for their fulfilment are captured in those dependencies. Figure 3-1 represents a design decision dependency diagram (D4), which aims to provide an overview of the relevant decision-making criteria and the relations between them. The grey nodes represent the categories with most dependencies (Building Site, Building Orientation, Building Envelope, Building Services, HVAC, Indoor Environment, Thermal Comfort, Energy Performance) and highlight not only the criticality of these decisions to the actual performance, but also the data that would be most relevant for goal-oriented analytics and knowledge discovery (Petrova et al., 2018). Understanding the links between the (critical) decision nodes and how they affect each other also has a significant impact on understanding the requirements to data collection, analysis and infrastructure needed to provide the evidence-based and context-aware decision support.

Furthermore, predictive models can be used in combination with the decision dependency network to quantify the weights of the dependencies, the criticality of the decisions, the variability of outcomes and the potential impacts. Understanding these contributes to the understanding of the data needs and the DSS features. Various kinds of data in different formats are generated during the entire building life cycle; however, not all knowledge discovery techniques are equally applicable to all types of data. To be able to determine the most appropriate analytical techniques to fulfil the knowledge discovery goals and close the gap between design and operation, the

following sections categorize the diverse building data types based on their phase of origin.

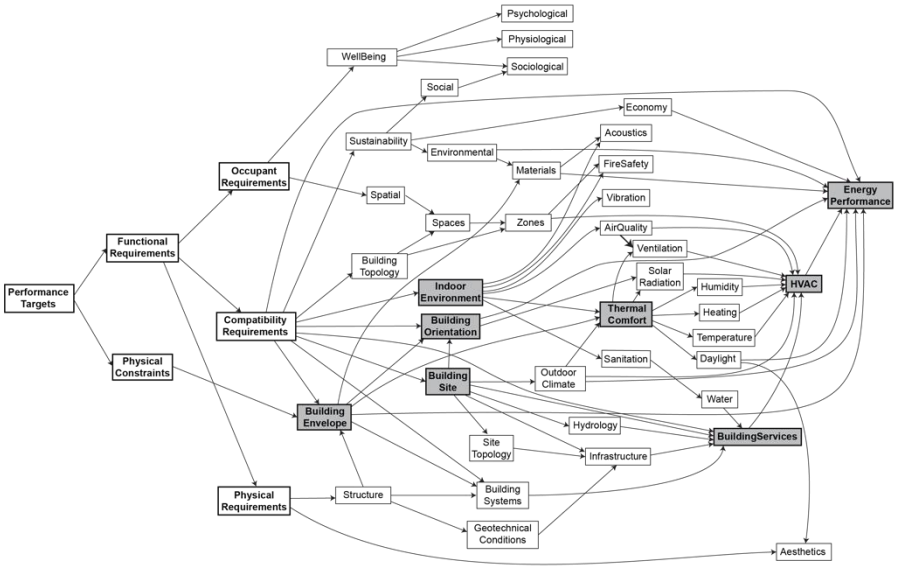

Figure 3-1: Design decision dependency diagram (D4 (Petrova et al., 2018)

3.1.2. DATA AND KNOWLEDGE IN THE BUILDING OPERATION PHASE

Operational building data from Supervisory Control and Data Acquisition (SCADA) systems and BMS is structured data usually represented in a two-dimensional tabular way. The columns in the tabular data represent the observed variables and the rows store the measurements with time stamps in accordance with the set measurement interval (Petrova et al., 2018). Han et al. (2012) state that the typical representations of operational building data allow to discover two main types of knowledge: static (cross-sectional) and dynamic (temporal). Static (cross-sectional) knowledge is discovered when each row of measurements is treated as an independent observation. In that case the temporal dependencies between the rows are ignored and the discovered knowledge mainly highlights relationships between the different observed variables. Cross-sectional knowledge discovery in operational building data can be used to identify interactions between system components, anomalies in operation, etc. (Han et al., 2012). On the other hand, Fan et al. (2015a) specify that dynamic knowledge discovery techniques consider both axes and is, therefore, highly useful for obtaining insights related to the dynamics of building operation. Such insights can be used to develop optimal control strategies, as well as fault detection and diagnosis. Also, temporal knowledge discovery allows to discover unsuspected patterns in data and the relationships between them.

Besides time stamps according to the set measurement interval, monitored building data usually includes energy consumption data (e.g. electricity consumption, heating, cooling loads, lighting, etc. [kW]), environmental data (e.g. temperature [C°], relative humidity [%], [CO_2], etc.), automation and control data (e.g. window opening and closing, shadings, etc. [0 or 1]), occupancy data, etc. These data change dynamically and influence building performance directly, which makes them a valuable input for knowledge discovery. Figure 3-2 represents dynamic parameters related to external conditions and operational data types typically collected from BMS (Petrova et al., 2018).

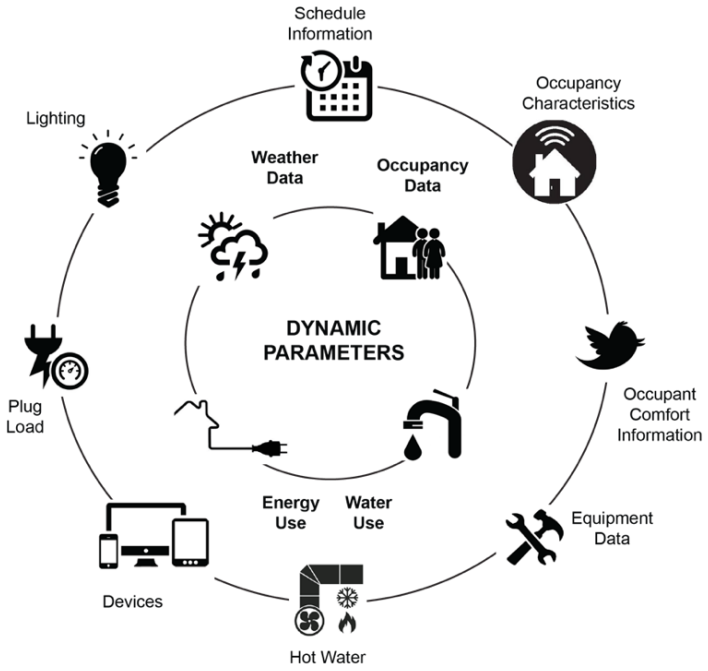

Figure 3-2: Dynamic building data parameters, based on taxonomy by Mantha et al. (2015) (Petrova et al., 2018)

3.1.3. DATA AND KNOWLEDGE IN THE BUILDING DESIGN PHASE

Data generation during the building design phase typically starts with design brief documentation and a conceptual BIM model, which later becomes the basis for development of various aspect and analytical models. The earliest stages usually aim for design space exploration in relation to design brief requirements and performance targets and include crucial choices related to building orientation, zoning, spatial arrangement, building materials, etc. Building geometry is one of the prominent data types at this stage, as it provides many of the main inputs required for simulation and performance analyses. With the development of the design, those parameters become

static and respond to the requirements and constraints as shown in the dependency diagram in Fig. 3-1. Figure 3-3 depicts the static building data parameters that to a large extent define the character of the building, also in terms of performance. Simulation data can also provide valuable insights into building performance that can inform future design. However, it has to be noted that such insights are more optimistic in comparison with the actual building performance, so model calibration with measured building data may boost their usefulness (Petrova et al., 2018).

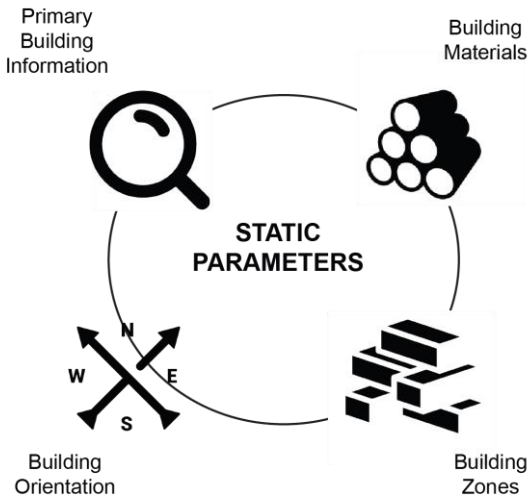

Figure 3-3: Static building data parameters, based on taxonomy by Mantha et al. (2015) (Petrova et al., 2018)

3.1.4. AN ANALYTICAL PERSPECTIVE ON BUILDING DATA

When it comes to the discovery of valuable insights in (building) data, of significant importance are the knowledge discovery goal and the suitability of the chosen analytical techniques. Understanding the data structure and representation is of vital importance to the understanding of the input needs of the data mining algorithms and, therefore, the effectiveness of the knowledge discovery process. To enable the selection of appropriate knowledge discovery techniques, the following list presents a definition of the different building data types from an analytical perspective (Petrova et al., 2018):

- Semantic design data: semantic data describing design features and their properties, including building elements, materials, object types, design brief data, etc.;
- Numeric geometric data: geometric data in a format optimized for geometric analysis;

- Binary geometric data: imagery and remote sensing data, such as point clouds;
- Numeric sensor data: real-time data streams from sensor networks acquired through SCADA and BMS;
- Numeric simulation data: data models containing simulation results.

Knowledge discovery in operational building data

Real-time monitored data usually updates continuously with data points to result in a data stream that gives an indication of the built environment's behaviour. Building operation is characterised by complex dynamics, caused by changes in the previously outlined dynamic parameters, i.e. changes in external conditions, occupant behaviour, systems utilization, etc., which usually do not co-occur simultaneously, but may exhibit a certain level of regularity (e.g. seasonal changes in weather, occupant behaviour, schedules of operation, etc.). Discovering dynamic dependencies in measured data is highly valuable and can positively influence decision-making related to choice of spatial design parameters, building components, control strategies, HVAC systems, etc. (Petrova et al., 2018).

In that relation, Fan et al. (2015a) state that temporal knowledge discovery can help capture relationships between monitored building variables over a particular time period. The authors ground that conclusion in the fact that operational building data is in essence multivariate time series data, where each observation is a vector of multiple measurements and control signals, and time intervals between the subsequent observations are fixed. As previously discussed in the state of the art review, several different approaches target knowledge discovery in time series data, i.e. events, clusters, motifs (frequent sequential patterns), discords (infrequent sequential patterns) and association rules (Fu, 2011).

To be able to effectively inform design decision-making in an evidence-based manner, it is important that the discovered knowledge increases the confidence of the decisions, while still allowing creativity and variability of design space exploration (Petrova et al., 2018). The objective of this research is to discover and reuse knowledge related to the dynamic behaviour of buildings and its influencing factors, which includes unsuspected patterns and the relationships governing them. Thus, the knowledge discovery and data mining approaches have to be carefully selected to fit that goal. According to Fu (2011), suitable for such an exploratory data analysis are motif discovery (frequent pattern mining) and ARM. Motifs are valuable elements of temporal knowledge discovery, because they allow to discover inherent regularities (or anomalies in case discords are targeted) in building operation and are valuable input for ARM. Important to consider here is the fact that frequent patterns in data do not necessarily start at the same time or have the same length, which makes motif discovery a highly useful approach, as it allows the exploration of such variations.

In addition, ARM can help discover associations between variables (Agrawal et al., 1993). Traditional ARM techniques usually targets cross-sectional knowledge discovery, but due to the complexity and dynamics of operational building data, the use of temporal ARM would be more useful, because it provides both an insight into the associations between the variables, as well as their temporal dependencies (Fournier-Viger et al., 2012). Such an approach to KDD would provide a solid foundation for evidence-based design decision support, as it can help identify complex building performance patterns over time and the dependencies in their occurrence.

Feature matching in geometric data

In line with the objectives of this research in terms of knowledge reuse, direct geometric pattern matching techniques can also be implemented and used to return results to a user query (Petrova et al., 2018). In that sense, several types of geometric data representations can be considered. A common example is IFC, which is a vendor-neutral data model aiming to capture building semantics and object properties along with 3D geometry in full detail (Borrmann et al., 2018).

Alternative open data models are also available, including the geometry ontology defined by Perzylo et al. (2015) and Well-Known Text (WKT)[11], which is a markup language that allows to specify geometry with simple strings based on common agreement. And while most WKT implementations refer to representation of 2D geometry in the geospatial domain, Pauwels et al. (2017) showcase that WKT can also be used for representation of 3D building geometry. Building geometry can also be represented using 3D mesh models. Yet, as stated in Petrova et al. (2018) such data is semantically less defined and direct geometric feature matching techniques are seldom useful in such case. The same applies to point cloud data, which is also used for geometry representation, but, similarly to 3D mesh models, such data presents limited semantics.

In terms of knowledge reuse, direct graph matching techniques can be used for semantically rich geometric data. SPARQL, CYPHER, and GraphQL are graph query languages, which can be used for graph matching in a CDE. Direct graph matching naturally requires the target geometric data to be represented in graphs, which can be the case for IFC, WKT, and geometric topology graphs.

Alternatively, as stated in Petrova et al. (2018), when geometric data is semantically less defined (point clouds and 3D mesh models), advanced geometric analysis algorithms can be applied, which aim at parsing input geometry and identifying characteristics. The extracted characteristics are typically semantic and can therefore reside in a semantic data structure. GeoSPARQL and BimSPARQL (Zhang et al.,

[11] http://www.opengeospatial.org/standards/wkt-crs.

2018b) are query languages that contain statements such as ‘within’ and ‘above’ and thereby allow to formulate geometric semantic queries.

Semantic queries

Direct semantic queries can also be used to retrieve information in response to domain-specific user requirements. Such queries can target a semantic integration layer over several repositories, semantic design data and/or attributes that may be inferred from knowledge discovery in operational building data or geometric feature recognition (Petrova et al., 2018).

As previously discussed in Chapter 2, domain specific ontologies are hereby of utmost value as they allow the representation of the different building data in a semantically well-defined and explicit way. However, as previously mentioned, an ontology-based approach may not be optimal when operational or geometric building data is targeted (Pauwels et al., 2017a). Ontologies can be used to capture static characteristics related to operational data and observed variables, such as averages, min–max values, features of interest, sensors, actuators, etc. Sensor data points, however, may significantly reduce efficiency if stored directly in the graph.

The results of the geometric analysis algorithms can be captured in semantic graphs through semantic annotations, but complete geometric matching would be most useful through the original data in a non-semantic format. A semantic integration layer can hereby help establish a connection between the semantic, the non-semantic numeric data (e.g. web server address of a sensor data warehouse), and geometric data (web server address of specific geometric data store). The purpose is to integrate the semantic, geometric and operational data, so that any system accessing the data can recognize the associations between the different data (Petrova et al., 2018).

3.2. HOLISTIC SUSTAINABLE BIM-BASED BUILDING DESIGN: PROPOSED FRAMEWORK AND SYSTEM ARCHITECTURE

As previously mentioned, design professionals approach decision-making in an iterative problem-solution manner, in which they devise solutions based on their background knowledge and by using their dedicated technology stack. A DSS implementation then has to be able to enhance human decision-making capabilities and not disregard the human for the sake of technological sophistication. Thus, effective decision support requires an in-depth understanding of the user needs. An insight into the cognitive processes occurring during design decision-making can provide valuable input for system design. The following section outlines the overall design thinking and decision-making processes and how they fit in a BIM-based process relying on knowledge reuse through a CDE, project data repositories and knowledge bases.

3.2.1. DESIGN THINKING AND PROBLEM SOLVING IN A DATA-DRIVEN DESIGN PROCESS

With each design iteration, design professionals explore a problem-solution space, thereby going through a continuous co-evolution of problem and solution (Maher & Poon, 1996; Dorst & Cross, 2001). As indicated in Petrova et al. (2018) the digital constituents of this process are typically stored in a CDE, which contains the multidisciplinary attributes of the design solutions as they come in sequentially. The design brief requirements and performance targets are hereby the main driver for the iterations and the decisions, and follow the continuous co-evolution of problem and solution. In that sense, both the interpretations of the requirements, as well as the solutions responding to those evolve throughout the phases of the design process. Ultimately, the design practitioners converge under the influence of the design brief and performance targets, which brings the team closer to a solution that fulfils both. The main objective design is to avoid widening of the cycles too much throughout the evolution towards convergence (Petrova et al., 2018).

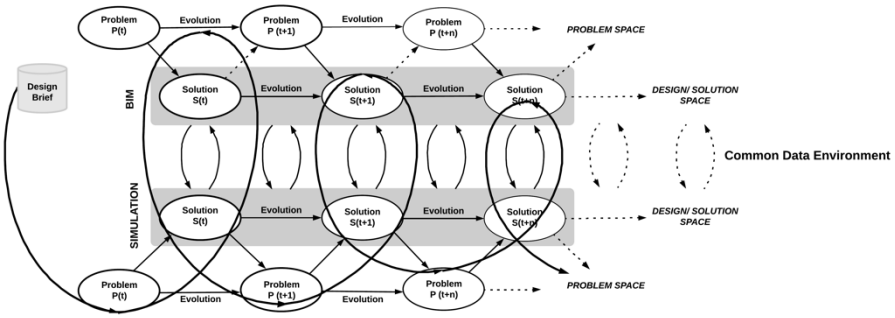

Figure 3-4: Problem-Solution cycles in collaborative building design (Petrova et al., 2018)

To give performance data and knowledge discovered in data a more prominent role in the design process described above, the way decision-makers utilise background knowledge needs to be influenced. This can be achieved by presenting the decision-maker with useful alternatives in the problem-solution space, which complement and build on the tacit knowledge in a structured way (Petrova et al., 2018).

3.2.2. LINKING DISCOVERED KNOWLEDGE, DATA AND BACKGROUND KNOWLEDGE

The proposed system architecture uses sensor data and various types of project data as an input for knowledge discovery. The top in Fig. 3-5 represents the active design environment, which communicates with the knowledge bases integrating various project data repositories (bottom in Fig. 3-5). Each project data repository collects all reference data linked together with the semantic integration layer. Important to note is that data is also kept in its native format. The main modules of the system architecture are outlined in the following sections.

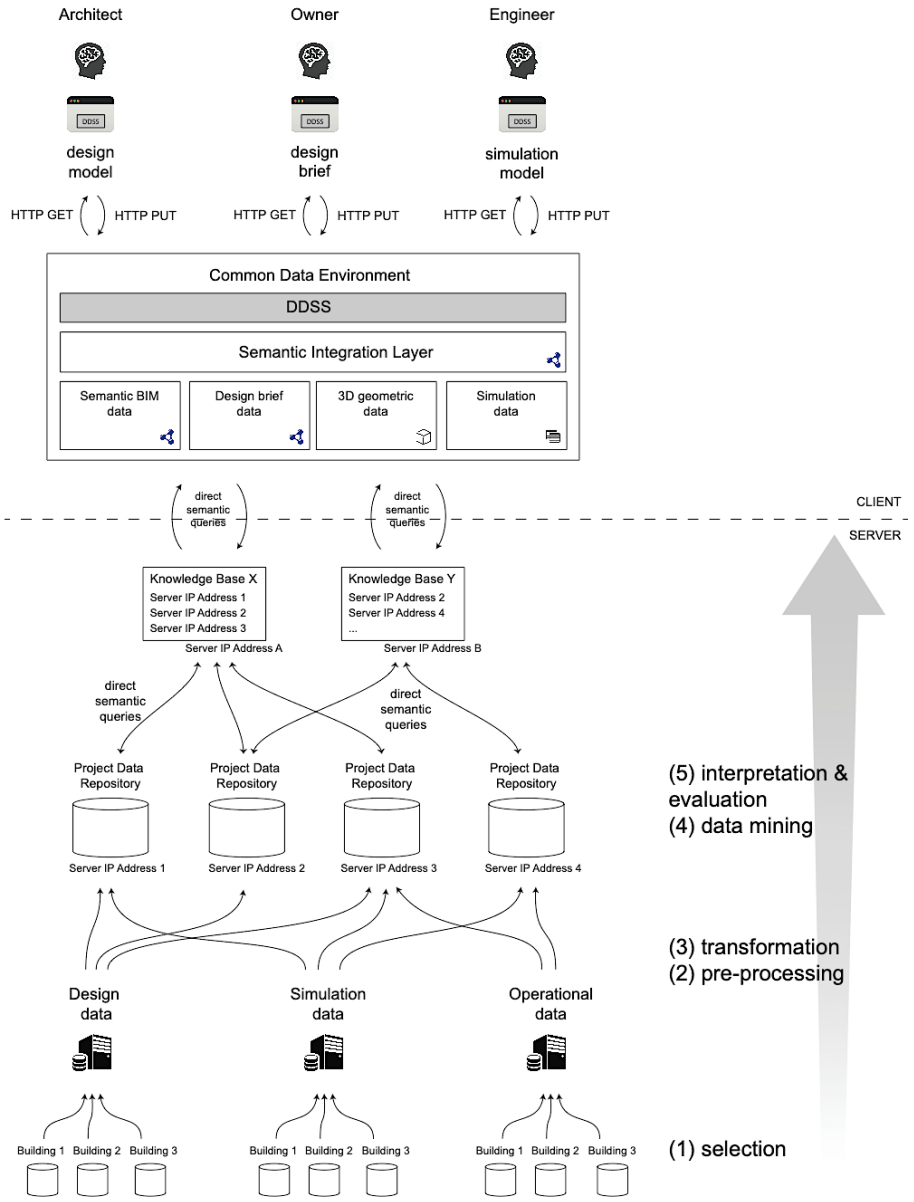

Figure 3-5: Proposed system architecture for evidence-based building design relying on knowledge bases

The active design environment

As stated in Petrova et al. (2018), even if a CDE is used in projects, it typically follows a file-based approach, which hinders the integrated view over the available information. State of the art initiatives aim at making the data available in an integrated manner with web technologies, which can also be used in the current context to make the CDE web-compliant and data-oriented, as opposed to the traditional document-based nature.

A system relying on web technologies is much more promising as it (1) enables semantic information retrieval and data management, (2) allows a larger diversity of knowledge discovery approaches, as data can be accessed and processed efficiently, while maintaining the same semantic identifiers, and (3) provides the necessary infrastructure for advanced semantic data mining techniques (Petrova et al., 2018). In such a setting, the web-based CDE is automatically filled with data using the HTTP protocol, which unburdens applications and users from having to store files on the server manually. In addition, data logging and versioning can be done in a much more efficient way. Considering that the purpose is to utilise data from multiple heterogeneous sources, the CDE would function optimally with a decentralized structure, which can be achieved using graph database approaches (Petrova et al., 2018).

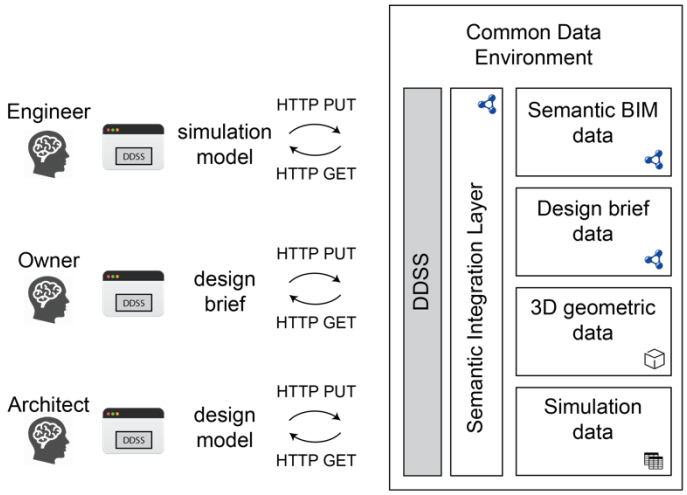

Figure 3-6: Integration of datasets in a web-based CDE (Petrova et al., 2018)

A graph-based approach is a prerequisite for the desired web of semantic building information and can serve as a backbone of the web-based CDE, thereby allowing to link the diverse datasets together, but also respect their original data structures.

The project data repository and semantic integration layer

As previously discussed, research shows that not all data can be efficiently maintained in a graph database or a triple store (Pauwels et al., 2017a). Thus, large volumes of numeric data (e.g. sensor data) are purposefully kept out of the semantic graph and stored in, for instance, SQL stores. Similarly, geometric data is maintained most efficiently in formats that can be parsed by geometric analysis algorithms (for geometric feature matching). To provide the necessary integrated view over the diverse datasets, a semantic integration layer is introduced, which has a thin, modular structure and maintains the links between the diverse datasets. The semantic integration layer captures the semantics of the heterogeneous data sources in a decentralized manner, while referring to the original data sources (Petrova et al., 2018).

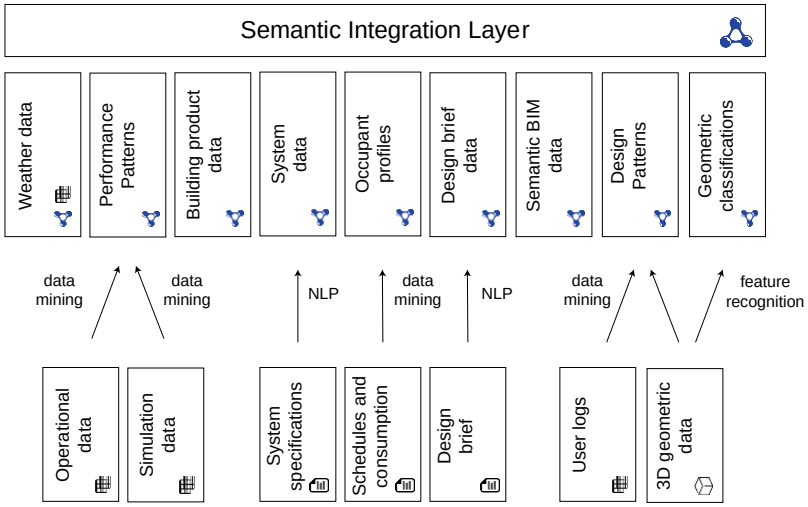

Figure 3-7: Overview of the project data repository with semantic integration layer

Maintaining this data structure instead of converting all data into linked data allows a much higher flexibility in terms of geometric feature matching and data mining. Ristoski & Paulheim (2016) indicate that a traditional data mining process can reside in a linked data context; however, this would disallow the use of many powerful feature matching and data mining algorithms that can be highly useful for knowledge discovery in geometric and operational building data. For that reason, semantic, geometric, and operational data are stored separately. In terms of retrieval, semantic queries alone cannot provide the same insight that can be obtained through data mining. However, relying solely on data mining approaches does not provide an integrated view over the diverse datasets. That also applies to geometric feature matching- relying only on geometric data to retrieve valuable knowledge from a project repository is not sufficient in a performance-oriented design setting.

Therefore, the diverse data have to be accessible and dynamically linked to allow information retrieval and design decision support on a holistic level. The semantic integration layer hereby provides the opportunity for integration of heterogeneous data and discovered building (performance) patterns, while still enabling semantic information retrieval through user-defined queries (Petrova et al., 2018).

Building a project data repository as proposed above requires several crucial steps and considerations. First, the reference data needs to be selected and transformed so that it fits the infrastructure of the project data repository. Preservation of data integrity is also an important topic and considerations in terms of data cloning and storing of local copies need to be made. Implementing a data selection process ensures that the data to be included in the project data repository for retrieval is in scope and the original data is also maintained secure. In a next step, the data can be cleansed and transformed in accordance with the needs of both the data mining algorithms and the structure of the project data repository (Petrova et al., 2018).

The knowledge base

The overarching fundamental element in the system architecture, which allows the retrieval of knowledge discovered in building data is the knowledge base. Each knowledge base can integrate multiple project data repositories enriched with performance patterns. Following the idea of decentralization, this research project assumes multiple knowledge bases, that can respond to different contexts, e.g. geographical, climate zone, building types, etc.

These knowledge bases are conceived as simple ‘registries’ or DNS-like servers. Each knowledge base consists only of a list of project data repositories and the IP-addresses of its servers. As such, the knowledge base is simply a collection of externally available servers. One can configure a new knowledge base, so that it collects project data repositories of relevance. In other words, the knowledge bases function as gateways or routing systems towards the project data repositories. In order for the overall information retrieval system to work, the queries that are sent to the knowledge bases need to be forwarded to the project data repositories (and responses need to be returned in the other direction), as displayed in Fig. 3-8.

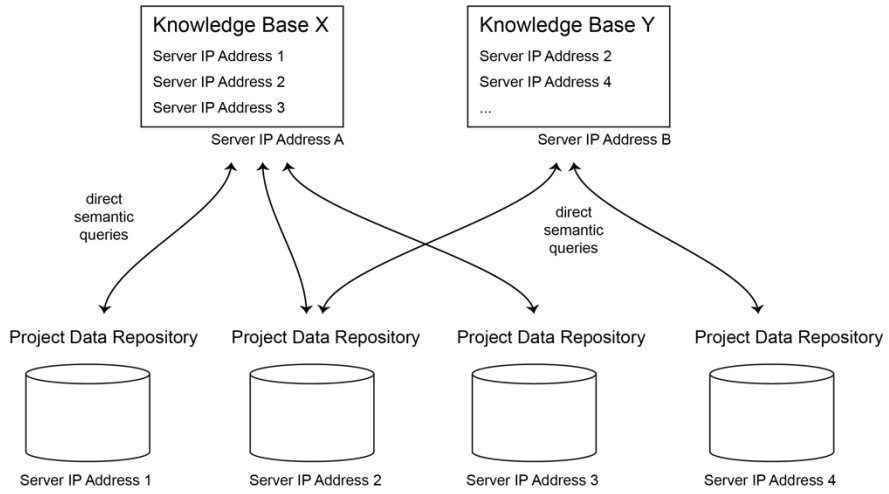

Figure 3-8: Knowledge bases integrating various project data repositories

3.3. TEMPORAL KNOWLEDGE DISCOVERY IN OPERATIONAL BUILDING DATA

The previous sections of this chapter outlined the types of building data, the knowledge that can be discovered and how they can impact the design process and the related decisions. The main conceptual framework for implementation of knowledge discovery, representation and retrieval was presented and the main modules were discussed. The following section presents the implementation of the knowledge discovery process and the related results. Knowledge discovery was performed according to the steps defined by Fayyad et al. (1996) and extended by Han et al. (2012). The implementation focuses on knowledge discovery in operational building data, therefore, motif discovery and ARM were used for extraction of insights from indoor environmental quality data collected from two use case buildings.

3.3.1. DATA MONITORING AND COLLECTION

Use case Gigantium: Public building with historical data and access to real-time data stream

Gigantium (34.000m²) is a cultural and sports centre located in Aalborg, Denmark. It opened in 1999 and has been renovated and extended multiple times over the last 20 years. It currently houses an ice skating arena, ice rink for training purposes, sports halls, a concert and exhibition hall, swimming pool and wellness areas, athletics and fitness hall, conference rooms, a cafe, and a visitors lobby. Operational data is collected through a network of 39 sensor nodes divided between the spaces. The

sensors measure temperature [°C], relative humidity (RH) [%], air pressure [hPa], Total Volatile Organic Compounds (TVOC) [ppb], CO_2 [ppm], illuminance [lux], motion and noise levels. The collected data is used for monitoring of indoor climate and thermal comfort levels for the visitors, facility management and providing information on space use. The collected sensor data is from the period 16.02.2018 to 17.05.2018 (Petrova et al., 2019; Petrova et al., 2018a).

Use case Home2020: Residential building with historical data and no access to real-time data stream

Home2020 (132m²) is a detached house near Aarhus, Denmark. It was completed in 2017 and rated as nearly zero energy building (NZEB) according to the Danish energy labelling standard. It hosts a kitchen, a master bedroom, a living room, three other rooms, two bathrooms, a utility room and a walk-in closet. The building occupants are a young working couple without children. District heating provides the heat supply to the building and is distributed to a floor heating system. The hot water and ventilation with heat recovery (85%) are provided by an air-to-water heat pump integrated in a compact unit. The ventilation system allows individual control of the air supply in the living room and bedrooms, and the extraction in the kitchen, bathrooms and utility room. The supplied air is adjusted according to the levels of CO_2 and relative humidity in each room. The house is also equipped with automatically controlled natural ventilation grids and skylights. The unit is running with a minimum airflow when the house is unoccupied and when a higher air supply is not required. The ventilation system is deactivated when the windows and doors are opened. External solar shading devices have been installed in the living room and bedroom and can be controlled automatically (Petrova et al., 2019).

A BMS is tracking several different parameters. Energy consumption is measured for district heating [MWh], floor heating pump [kWh], ventilation system [kWh], control system [kWh], and kitchen appliances [kWh]. Measurements for the compact unit include outdoor air temperature [°C], return air temperature [°C] , return air relative humidity [%], hot water temperature [°C], supply air temperature [°C], heat pump temperature [°C], ventilation speed [steps]. Both hot and cold water consumption [m^3] are also tracked. In terms of indoor environmental quality, sensors register temperature [°C], CO_2 [ppm], relative humidity [%], and damper opening [min/ max]. The data is collected with a measurement interval of five minutes and the used dataset for the period 01.12.2017 to 31.10.2018 (Petrova et al., 2019).

3.3.2. DATA PREPARATION AND CLEANSING

In the case of Gigantium, all data is collected in a relational database behind an open data visualization and monitoring platform (Grafana)[12]. The combination of a database and GUI dashboard interface allows real-time data monitoring and acquisition on

[12] https://grafana.com/

demand. As no direct live access to the database was available at the point of the experiment, the data mining preparation required a number of CSV exports from the monitoring platform. In the case of Home2020, no database access or management system with GUI is available, and the raw data is acquired in CSV files (data logs), with each CSV file containing the sensor data from one day (a total of 335 log files, each containing measurements of 76 observed variables with a five-minute measurement interval). For the knowledge discovery round in this research, only indoor environmental quality data is selected from the full datasets from both use cases, i.e. temperature [°C], CO_2 [ppm], and relative humidity [%] for both Gigantium and Home2020, and air additionally pressure [hPa] and Total Volatile Organic Compounds (TVOC) [ppb] for Gigantium.

After the data selection, the implementation of the KDD steps proceeds with cleansing and preparation of the data according to the need of the selected mining algorithms. The data from Home2020 does not contain any missing values or noise, as the dataset had either already been treated to remove inconsistencies or the quality was rather high. Therefore, in the cleansing and preparation round, only the sensor data logs for the period between 26-31.10.2018 are discarded, as they contain measurements of new observed variables that have been added to the logs as a result of a newly implemented automation and control strategy. As data mining results from only five days would not have any statistical significance in terms of building behaviour, these logs are excluded from consideration (Petrova et al., 2019).

In the Gigantium use case, however, several inconsistencies and missing data points were discovered, mostly due to downtime of some of the sensor nodes. That also includes nodes that have been inactive during the entire three-month timespan or such that started recording measurements considerably later. Furthermore, initial screening identified several outliers, e.g. room temperature values over 400°C, which are clearly erroneous. Missing data fields and removal of null values is performed with five iterations of multiple imputation by running the Expectation Maximisation bootstrap algorithm using the tool Amelia[13] in R. Outlier detection and removal is also performed. Furthermore, the sensor data is classified on a per sensor node, per room and per observed variable basis, to allow more dedicated analyses (Petrova et al., 2019; Petrova et al., 2018a).

In preparing for the next step in the KDD process (data mining), all data is loaded into a locally created Java code library containing Measurement classes, with each Measurement containing a Datetime stamp and a set of Property values. Each Property value records the type of observation and its value, together with a number of additional metadata. After the necessary preparatory steps, 94.434 measurements in total are parsed and loaded for the Home2020 case (Petrova et al., 2019).

[13] https://gking.harvard.edu/amelia

3.3.3. TRANSFORMING TIME SERIES DATA INTO SYMBOLIC REPRESENTATIONS

To prepare the data for frequent pattern discovery and ARM, Symbolic Aggregate Approximation (SAX) is applied on the loaded measurement values, both in the Gigantium and Home2020 case. As defined by Lin et al. (2007), SAX allows for dimensionality reduction and indexing with a lower bounding distance measure.In other words, SAX allows to reduce a large dataset to a smaller one, without losing the fidelity and characteristics of the data. To reduce the time series data from *n* dimensions to *w* dimensions, the data is divided into *w* segments, and each segment is replaced by the average of its data points (Piecewise Approximate Aggregation (PAA)). The value of each segment is then replaced by a symbol (Lin et al., 2007). Important to note here is that deciding on the number of SAX symbols and segments is essentially a task for the data analyst, and, therefore, it can potentially affect the results. The resulting symbolic representations of time series data allow using various machine learning algorithms for effective motif discovery and anomaly detection.

In the case of Home2020, this transformation step was done using the same Java library that was created for this experiment in the pre-processing step, combined with the SPMF open-source data mining library[14]. As indicated in Petrova et al. (2019), 7.869 segments were retrieved for Home2020, which implies hourly SAX representations (one symbol per hour representing the average of all 12 measurements per hour as a result of the measurement interval). Seven SAX symbols (1-7) were decided on for the SAX transformation of the dataset based on screening of the general behaviour of all observed variables in all rooms. Based on an analysis of the difference between minimum and maximum values of the observations, seven was selected as a number of symbols that would create intervals that fit the variances in the measured values of all observed variables in all rooms (Petrova et al., 2019).

As a result of the SAX representation, the complete sequence of data points is replaced by a symbolic representation such as 32222223222222223333..., with each SAX symbol representing an interval of data values (e.g. 2 = [22.86950723073572, 23.704365409749624]). Figure 3-9 presents an example of the seven SAX symbols and their corresponding values for the temperature observations in the bedroom (Petrova et al., 2019).

[14] http://www.philippe-fournier-viger.com/spmf/

```
1 [-Infinity,22.86950723073572]
2 [22.86950723073572,23.704365409749624]
3 [23.704365409749624,24.355554789380466]
4 [24.355554789380466,24.956652678270476]
5 [24.956652678270476,25.60784205790132]
6 [25.60784205790132,26.442700236915222]
7 [26.442700236915222,Infinity])
```

Figure 3-9: SAX symbols corresponding to temperature values in the bedroom in Home2020 (Petrova et al., 2019)

After replacing each sensor data point with its symbolic representation, co-occurrence matrices representing the co-occurrence of SAX symbols are computed on a per-month basis. An excerpt from a resulting co-occurrence matrix of SAX representations is shown below (Petrova et al., 2019):

Temperature_Bedroom	3 2 2 2 2 2 2 …
CO_2_Bedroom	4 3 3 1 3 7 7 …
RH_Bedroom	3 3 4 7 7 5 5 …
Temperature_LivingRoom	2 2 2 2 2 2 2 …
CO_2_LivingRoom	6 5 5 2 3 6 7 …
RH_LivingRoom	3 3 4 7 7 7 6 …

3.3.4. MOTIF DISCOVERY

After obtaining the sequences of SAX symbols, frequent repetitive patterns or 'motifs' can be mined. This is done by identifying the Longest Repeated Substrings (LRS) within each sequence of SAX symbols with a custom implementation of the Suffix Tree algorithm (Weiner, 1973; Ukkonen, 1995), which is highly effective in combinatorial pattern matching efforts. The algorithm serves to find the longest repeated substring within a string, yet, in this case, an altered implementation was made, which also outputs all other repeated substrings from input SAX sequences. This is done for both cases on a per month/per room/per observed variable basis (e.g. LRS output for CO_2 representations in the bedroom in August) (Petrova et al., 2019). An example of a set of motifs discovered in SAX sequences for a particular observed variable is presented in Fig. 3-10. The output includes the patterns of SAX symbols, the number of times they appears for each month- observed variable- room sequence, and the index in the sequence where the pattern starts.

```
345555 - 3 - 13;103;130;
444333 - 5 - 78;167;196;504;559;
4445555 - 4 - 29;178;244;642;
44544 - 3 - 124;222;241;
455555556 - 3 - 14;246;598;
455556 - 4 - 31;131;180;644;
54433 - 4 - 62;224;363;432;
55544 - 6 - 107;160;191;217;361;636;
555666 - 10 - 133;147;182;251;309;382;603;621;646;690;
6555554 - 3 - 157;188;723;
66655 - 8 - 141;155;186;301;428;629;681;706;
6667666 - 3 - 137;297;386;
```

Figure 3-10: A set of LRS found in the SAX sequences (Petrova et al., 2019)

Of course, not all repeated substrings are equally valuable and a number of those need to be discarded. This effort aims to identify only disjoint, non-redundant and non-overlapping patterns, therefore, overlapping and redundant patterns are excluded. This is done through a manual evaluation step, in which all output patterns are evaluated based on length, frequency and evolutionary character as criteria for "interestingness". This results in a final set of 27 files containing numerous discovered motifs per room, observed variable and month for Home2020 and 14 discovered motifs for Gigantium (Petrova et al., 2019; Petrova et al., 2018a).

In both cases, the resulting motifs are used to compute the co-occurrence matrices that show which motifs co-occur at any moment in time. Both for the Gigantium and Home2020 cases, co-occurrence matrices were computed to track co-occurrences of motifs in all observed variables (i.e. temperature, relative humidity, etc.). Figure 3-11 shows a visualization of the motifs discovered in the SAX representations of the sensor data from the visitors' café in Gigantium and their co-occurrences.

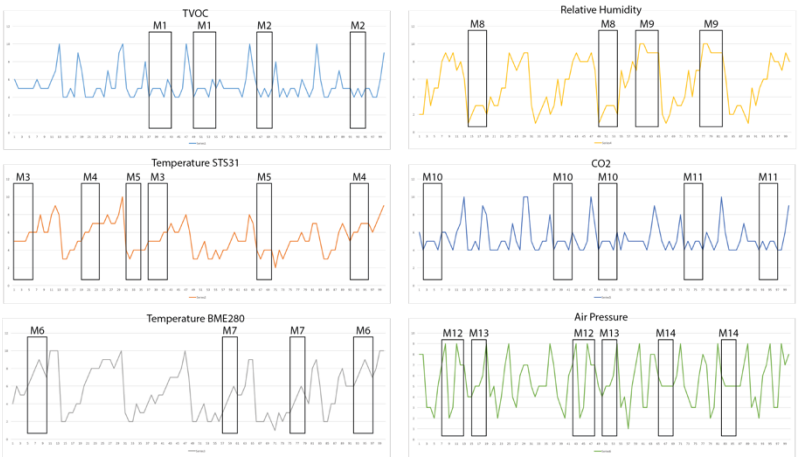

Figure 3-11: Co-occurring motifs discovered in the indoor environmental quality data measurements from the cafe in Gigantium (Petrova et al., 2019; Petrova et al., 2018a)

In the Home2020 use case, however, the significant number of discovered motifs (nine co-occurrence matrices of 730 columns each) resulting from the much larger dataset requires a structured automated approach to achieve the identification of motif co-occurrences. Therefore, in a first round, the motifs are assigned IDs and visualized in heatmaps (Fig. 3-12) to ease the detection of the co-occurrences and understand the pattern distribution throughout the sequences. Furthermore, as described in Petrova et al. (2019), the co-occurrences are calculated and composed in memory using the pattern IDs, thereby taking into account that multiple patterns may occur within the same sequence of SAX symbols. Each matrix is "stepped through" one Datetime value at a time and each time two or more patterns co-occur, a co-occurrence object is created.

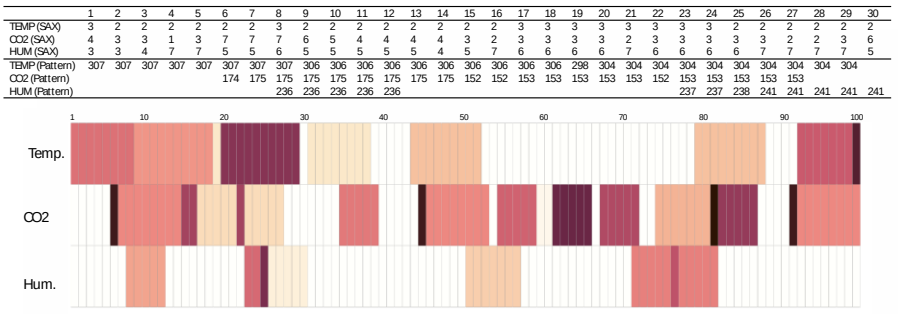

Figure 3-12: A heatmap visualization of the motif co-occurrences, their IDs and corresponding SAX representations in the sequence of January for the bedroom in Home2020 (Petrova et al., 2019)

For each co-occurrence, density of the co-occurrence is computed and traced. If a co-occurrence consists of two motifs, density of co-occurrence can be either 1 or 2 (overlap of 50% in one or two directions, respectively); if a co-occurrence consists of three motifs, density of co-occurrence can be anything from 3 to 6 (overlap of 50% between all three of the included patterns). This continues as the co-occurrences consist of more than three co-occurring patterns. The final co-occurrences for each room-variable-month combination are the starting point for ARM. The detailed computational method for the co-occurrence matrices can be found in Petrova et al. (2019).

3.3.5. ASSOCIATION RULE MINING

As a final step, ARM is performed, starting from the bags (multisets) of co-occurrences. Each co-occurrence is considered a 'transaction' and the totality of all transactions constitutes the 'transaction database' required for the rule mining. With this transaction database, it is possible to use the SPMF data mining library again. ARM is performed with an implementation of the FP-growth algorithm in SPMF. The output of the algorithm consists of the targeted association rules, including the measures of "interestingness": support and confidence, which indicate how frequently

a rule appears in the data and how often it is found to be true respectively (Agrawal, 1993).

Several hundred association rules are discovered for the Home2020 case, whereas the Gigantium case resulted in significantly fewer association rules, most likely because of the smaller dataset. Figure 3-13 shows an excerpt of the list of association rules discovered from the data from the living room in Home2020 in the month of August (Petrova et al., 2019).

```
452 ==> 489 #SUP: 1 #CONF: 1.0
453 ==> 485 #SUP: 3 #CONF: 0.6
454 ==> 481 #SUP: 1 #CONF: 0.5
456 ==> 484 #SUP: 2 #CONF: 0.6666666666666666
457 ==> 488 #SUP: 1 #CONF: 1.0
459 ==> 481 #SUP: 1 #CONF: 0.5
459 ==> 488 #SUP: 1 #CONF: 0.5
482 ==> 460 #SUP: 1 #CONF: 0.5
460 ==> 482 #SUP: 1 #CONF: 0.5
460 ==> 485 #SUP: 1 #CONF: 0.5
457 488 ==> 378 #SUP: 1 #CONF: 1.0
378 488 ==> 457 #SUP: 1 #CONF: 0.5
378 457 ==> 488 #SUP: 1 #CONF: 1.0
457 ==> 378 488 #SUP: 1 #CONF: 1.0
459 488 ==> 378 #SUP: 1 #CONF: 1.0
378 488 ==> 459 #SUP: 1 #CONF: 0.5
378 459 ==> 488 #SUP: 1 #CONF: 1.0
459 ==> 378 488 #SUP: 1 #CONF: 0.5
```

Figure 3-13: A part of the association rules obtained for the living room in August in Home2020 (Petrova et al., 2019)

Important to note here is that not all of the discovered association rules will be interesting or present novel insights. Further evaluations are required to discover the rules with the highest "rule surprisingness" level. Such an evaluation may include considerations related to the combined effect of the support and confidence measures or a domain expert assessment to identify the strong and interesting rules potentially indicating novel insights related to building performance (Petrova et al.,2019).

Despite the fact that a plethora of rules were discovered, at this point they represent merely a statistical output, which is the result of the knowledge discovery process. To become useful to an end user through a holistic evidence-based decision support mechanism, the discovered output needs to be represented in a format suitable for retrieval and meaningful to both machines and human users.

3.4. KNOWLEDGE REPRESENTATION AND RETRIEVAL FROM THE KNOWLEDGE BASE

Regardless of the interestingness of the retrieved motifs and association rules, the end user (designer, engineer, architect, etc.) has not benefitted from the knowledge discovered in the data yet, as all of the motifs and rules have not reached the end user yet. Therefore, according to the suggested system architecture and the introductory sections, the relevant building data and discovered knowledge need to be made accessible to the end users to enable evidence- based design decision support. Ideally, each type of data and knowledge discovered in data is stored in the most suitable possible format and the different datasets are linked across domains (semantic integration layer).

As previously stated, semantic web and linked data techniques allow to represent and link datasets together, therefore, these technologies are a natural choice for representation of the motifs and association rules if they need to be integrated with other available data. However, some of the limitations found in the literature need to be taken into account, i.e., the swollen graph issue, the stability of ontologies, and live data generation. The following section indicates both how linked data technologies are used to represent and retrieve the discovered knowledge, and how these limitations may be overcome.

3.4.1. SEMANTIC REPRESENTATION OF BUILDING DATA AND PERFORMANCE PATTERNS

Both the Gigantium and Home2020 buildings were modelled using the LBD ontologies and modelling principles (see Chapter 2). This includes the namespaces and prefixes listed in Fig. 3-14:

```
@prefix seas: <https://w3id.org/seas/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix bot: <https://w3id.org/bot#> .
@prefix geo-ext: <http://eapetrova.com/voc/geoextension#> .
@prefix bmeta: <http://eapetrova.com/voc/buildingmetadata#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix ssn: <http://www.w3.org/ns/ssn/> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix om: <http://www.ontology-of-units-of-measure.org/resource/
  om-2/> .
@prefix ptn: <http://eapetrova.com/pattern/> .
@prefix list: <https://w3id.org/list#> .
@prefix inst: <https://home2020.dk/instances#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
```

Figure 3-14: Namespaces used in the RDF graph (Petrova et al., 2019)

Both Home2020 and the Gigantium building have been modelled as RDF graphs according to the BOT ontology. These graphs contain the description of the buildings,

building storeys and spaces. The latitude, longitude, and altitude of the building are also included using geospatial ontologies, as well as an OpenStreetMap (OSM) location[15]. As 3D geometry or BIM models are not available in either of either cases, geometry is included at a bare minimum, using WKT strings for 2D space boundary representations, and thus also leaving out geometric feature recognition as an information retrieval option. A part of the resulting graph in the case of Home2020 is presented in Fig. 3-15.

Sensor nodes are included in the graph using the SSN and BOT ontologies. The ssn:hasProperty predicate links the spaces to the sensor observations that are made by the nodes hosted inside. Furthermore, the bot:containsElement containment relation links each space to the sensor node that it contains, and each node is linked to each individual sensor and the sensor observations it produces. The SOSA and OM (Units of Measure) ontologies are further used to include numerical measures, datetime of , measurements and units for each observation (Petrova et al., 2019).

```
inst:Home2020BuildingSite
  rdf:type owl:NamedIndividual, bot:Site ;
  rdfs:label "Site of the building"@en ;
  bot:hasBuilding inst:BuildingHome2020 .

inst:GroundFloor
  rdf:type owl:NamedIndividual, bot:Storey ;
  rdfs:label "Ground floor of the building"@en .

inst:BuildingHome2020
  rdf:type owl:NamedIndividual, bot:Building;
  rdfs:label "Passive house"@en;
  bot:hasStorey inst:GroundFloor ;
  bot:hasSpace inst:Kitchen , inst:LivingRoom , inst:Bedroom ;
  geo:lat "56.0914290" ;
  geo:long "9.7958060" ;
  geo:alt "16" ;
  geo-ext:inOSMLocation <https://www.openstreetmap.org/node
    /3721416569> .

inst:Kitchen
  rdf:type bot:Space, sosa:FeatureOfInterest ;
  bot:containsElement inst:sensorNode_1 ;
  rdfs:label "Kitchen"^^xsd:string ;
  ssn:hasProperty inst:Kitchen-CO2, inst:Kitchen-Temperature, inst:
    Kitchen-Humidity .
```

Figure 3-15: A snippet of the RDF graph of Home2020 (Petrova et al., 2019)

As such, the Gigantium and Home2020 buildings are represented as much as possible according to best practices defined by the LBD community group and

[15] https://www.openstreetmap.org/

recommendations presented in diverse research initiatives (see Chapter 2). Essentially, the creation of new ontologies is kept to a bare minimum as the purpose is to reuse existing ontologies. As one of the main objectives is to be able to use the knowledge discovered in the previous step, a "pattern" ontology (:ptn) was built for the purpose. It enables the representation of the discovered association rules, including their ptn:confidence, ptn:absoluteSupport, and ptn:relativeSupport measures. The association rules (inst:associationRule 1) are linked to the sensor nodes they originate from using ptn:hasAssociationRule predicates. Furthermore, the association rules link to ordered lists of motifs on the left-hand side (ptn:LHS) and right-hand side (ptn:LHS) of each rule (see Fig. 3-13). The motifs are represented with their correspondingspace, observed variable and SAX symbols with the lower and upper bounds of the interval (Fig. 3-16) (Petrova et al., 2019).

```
inst:associationRule_1
  rdf:type ptn:AssocationRule ;
  ptn:LHS (inst:Motif_45) ;
  ptn:RHS (inst:Motif_137) ;
  ptn:confidence "0.5"^^xsd:double ;
  ptn:absoluteSupport "1"^^xsd:double ;
  ptn:relativeSupport "0.5"^^xsd:double .

inst:motif_45
  rdf:type ptn:Motif ;
  ptn:SAXsequence "11122"^^xsd:string ;
  ptn:space inst:Kitchen ;
  ptn:month "8"^^xsd:string ;
  ptn:SAXsequenceFull (inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7
    b177e237bc0 inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7b177e237bc0
     inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7b177e237bc0 inst:
    SAXSymbol_41fadfdb-6560-4e96-9a7f-bc405f453452 inst:
    SAXSymbol_41fadfdb-6560-4e96-9a7f-bc405f453452 );
  ptn:observedVariable "CO2"^^xsd:string .

inst:SAXSymbol_36ef82d8-57c9-4e0a-a0bc-c1c66404b02b
  rdf:type ptn:SAXSymbol ;
  ptn:symbol "5"^^xsd:int ;
  ptn:lowerBound "645.651281059915"^^xsd:double ;
  ptn:upperBound "700.959674546294"^^xsd:double .
```

Figure 3-16: A snippet of the RDF graph of Home2020 with motifs and associated rules modelled according to the built for the purpose PATTERN ontology (Petrova et al., 2019)

In terms of adding sensor data, a key decision needs to be made in terms of adding the data to the graph or not. As indicated in the state of the art, sensor data can be added directly to the graph, in the form of RDF triples, thereby relying on the SOSA and SSN ontologies. This approach was taken for the Home2020 building. However, this results in a considerably bigger graph, which reduces query performance and ease of use (Petrova et al., 2019). Moreover, sensor data is currently usually not retrieved and used as RDF graphs in this specific domain; many more algorithms and tools are oriented towards tabular sensor data. Therefore, the Gigantium case explored the inclusion of a URL in the graph, which points to the relevant sensor data in the original SQL store behind the Grafana API (Fig. 3-17). As indicated in Petrova et al. (2019),

a custom datatype property points to a web address that returns the data values as requested using the HTTP protocol. It is possible to add attributes to the HTTP requests, thereby setting query parameters such as time frame and refresh rate (e.g. from=now-30d&to=now&refresh=30s). The result includes the pointer to the data stream for a sosa:Result of a sosa:Observation.

```
inst:room_1
    rdf:type bot:Space ;
    rdfs:label "Main hall" ;
    bot:hasSpace inst:room_2 ;
    bot:containsElement inst:sensorNode_00000097, inst:
      sensorNode_000000B0, inst:sensorNode_00000077 ;
    geom:hasGeometry "2000, 3000, 4000, 6000"^^wkt:linestring.

inst:sensorNode_00000097
    rdf:type sosa:Platform, bot:Element ;
    rdfs:label "00000097" ;
    bmeta:observation "Space use" ;
    sosa:hosts inst:sensor_00000097_1 ;
    bmeta:placement "Placed in the middle of the hall, 8m above the
      floor."


inst:result_1
    rdf:type sosa:Result ;
    rdfs:label "Result of observation of Relative Humidity" ;
    bmeta:values "https://gigantium.dk/Gigantium2018instances?orgId
      =1&datastream=true" .
```

Figure 3-17: A snippet from the RDF graph of the Gigantium use case building (Petrova et al., 2019)

When taking the second approach, a number of caveats need to be taken into account. The application that consumes the data needs to be configured or implemented so that it expects these URLs and knows what to do with it. This requires additional programming to retrieve the values and display them in a GUI. Furthermore, this storage method requires the API of the original SQL-based database to be stable.

3.4.2. PROJECT DATA REPOSITORY AND KNOWLEDGE BASE

To achieve optimal information retrieval results for design decision support, the information retrieval should exploit a rich knowledge base hosting heterogeneous data and discovered knowledge from diverse buildings. The Gigantium and Home2020 cases serve as excellent examples for testing the overall data modelling approaches. Next, larger scale data repositories are needed that rely on the same data modelling

approach. Such heterogeneous knowledge bases are vital to the performance of the intended decision support mechanism.

By absence of such openly available data repository, a new knowledge base relying on distributed project data repositories was built as part of this research endeavour. This data repository consists of a self-owned collection of 531 building models originally available in the IFC data model. The models are converted to linked data by the use of the IFC-to-LBD converter[16]. The resulting RDF graph and the contained data are compliant with the overall LBD approach, which makes them easy to query using the SPARQL query language.

As described in Petrova et al. (2018a) and Petrova et al. (2019), the conversion results in a collection of two Stardog triple stores, containing a total of 36 Million triples divided between them. The purpose of spreading the data over two stores is to create a scenario close to the real world, where more than one repository is available and retrieval happens through a federated query approach. The data includes 372 bot:Building instances, 3,523 bot:Zone instances, 2,117 bot:Space instances, and 615,452 bot:Element instances. The bot:Element instances also include a product type (wall, window, etc). The graphs for the Home2020 and Gigantium use case buildings are added to this repository, including the sensor data, discovered motifs, association rules, etc (Petrova et al., 2019; Petrova et al., 2018a; Petrova et al., 2018b).

The created knowledge base is a proof of concept for the backbone of the outlined system architecture, namely a set of distributed knowledge graphs of diverse buildings, which can be further enriched with product data, design requirements data, geometric data, geospatial data, etc. This knowledge base is distributed over multiple databases, and thus allows to mimic the desired decentralised knowledge base in the targeted system architecture. Further repositories could be made available in the future, as owners make their building data (openly) available.

3.4.3. INFORMATION RETRIEVAL

As indicated in the introductory sections, information retrieval needs to be triggered from within the design environment used by the design team in order for the system to be user-centred. Considering the overall impact of BIM tools and approaches, this design environment will most often be a BIM tool, making the targeted system BIM-based and user-centred. The discovered motifs and association rules can be used to inform design decisions related to spatial design, thermal comfort, indoor climate, HVAC system design, etc. In order to obtain reference knowledge from the building data repository, SPARQL queries will be executed depending on the context of the design team and the current project (Petrova et al., 2019).

[16] https://github.com/jyrkioraskari/IFCtoLBD

Petrova et al. (2018b) discusses such a setup with an active design case example. Yet, it needs to be mentioned that the scope of this research project also includes, a recommender system setup, and much of the information retrieval functionality will rely on the way in which the recommendations are made (Chapter 5). As a result, this final part of Chapter 3 limits to indicating which query functionality is currently available for the knowledge base.

One of the targeted use cases is a design team working in a BIM environment, which would benefit from relevant knowledge discovered from previous building projects and actively used buildings. In such a case, a key query would be to retrieve buildings or spaces of the same type. For such buildings or spaces, various evaluations can be made, for instance, in terms of indoor environmental quality by the discovered patterns, and the embedded in the RDF knowledge graph data on systems, materials, building components, etc. such a dedicated retrieval can help decision-making in terms of thermal comfort, daylight, HVAC system design, etc. in a given context.

For such a case, the rdfs:label tags can be used to retrieve buildings and spaces of a particular type. These tags are currently unstandardized strings. It would be better if all buildings had the same standardised classification tags used throughout the repository (e.g. Getty AAT tags[17]) (Petrova et al., 2019). Alternative queries to obtain reference buildings and/or spaces are of course also possible. Figure 3-18 shows an example SPARQL query, which retrieves a list of relevant building and space URIs. This is a federated query, relying on the SERVICE and UNION keywords in SPARQL to be able to query both building data repositories at once, thereby fulfilling the knowledge base vision (Petrova et al., 2019; Petrova et al, 2018b).

[17] http://www.getty.edu/research/tools/vocabularies/aat/

```
SELECT ?b WHERE{
    {
        SELECT ?b WHERE {
            SERVICE <http://localhost:5820/BuildingDataRepo1/query>
                {
                    ?b a bot:Building .
                    ?b bot:hasSpace ?s .
                    ?s rdfs:label "Kitchen"^^xsd:string ;
                }
        }
    }
    UNION
    {
        SELECT ?b WHERE {
            SERVICE <http://localhost:5820/BuildingDataRepo2/query>
                {
                    ?b a bot:Building .
                    ?b bot:hasSpace ?s .
                    ?s rdfs:label "Kitchen"^^xsd:string ;
                }
        }
    }
}
```

Figure 3-18: SPARQL query for relevant buildings, federated over the distributed Stardog project data repositories constituting the knowledge base (Petrova et al., 2019)

The returned URIs make available the data and the relevant knowledge discovered for that particular building. Figure 3-19 shows an example with Home2020 as a returned result responding to the query for buildings with spaces of type “kitchen”. The top of the figure shows the BOT topology of the building with the kitchen hosting the sensor nodes and the discovered performance patterns and association rules (red and yellow nodes respectively). Moreover, the three contained sensors (CO2, Temperature, Relative Humidity) can be retrieved, including the actual observation measurements and units (bottom and left, purple, lime green and cyan nodes) (Petrova et al., 2019).

As stated in Petrova et al. (2019), the returned URIs serve as reference points for further retrieval of additional knowledge. For instance, these URIs can be used in the BIM environment for further retrieval and evaluation of the performance patterns and rules associated with the retrieved spaces. Figure 3-20 shows an example of such a second round query, targeting specifically observations, motifs and rules.

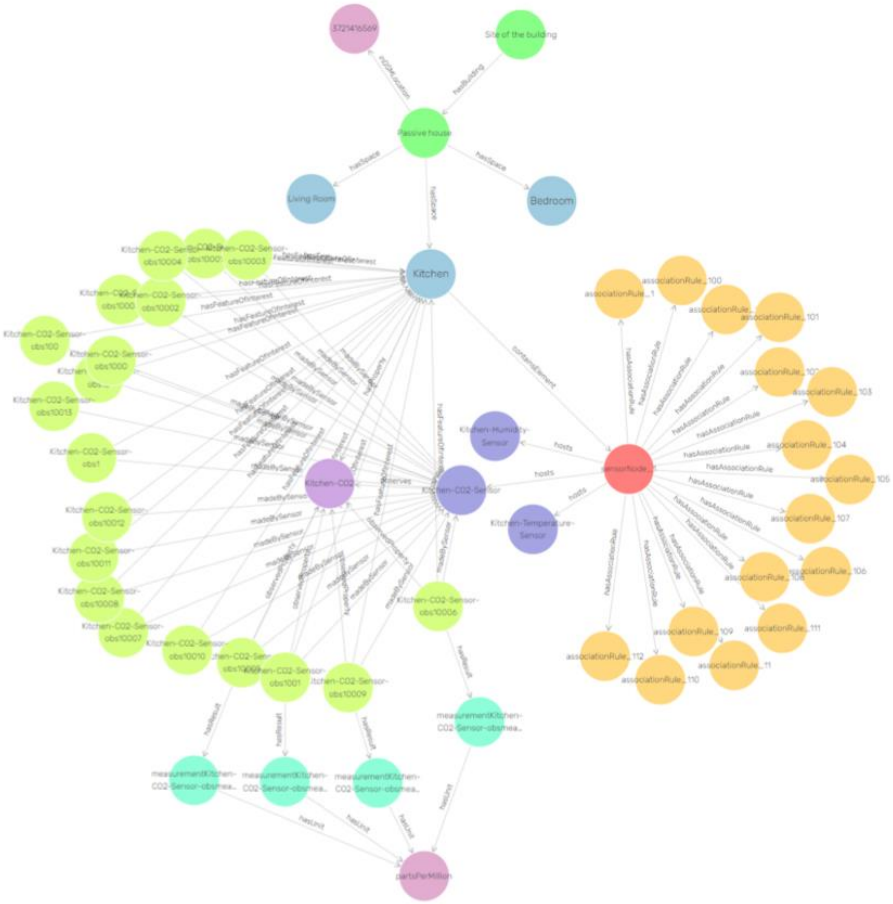

Figure 3-19: A resulting semantic graph in response to the executed query containing the building URIs and the related spaces, sensor nodes, sensor data and the motifs and association rules discovered in the data (Petrova et al., 2019)

```
SELECT ?sensor ?ar ?obs WHERE {
  ?s a bot:Space .
  ?s bot:containsElement ?sn .
  ?sn sosa:hosts ?sensor .
  ?sn ptn:hasAssociationRule ?ar .
  ?sensor ssn:observes ?obsp .
    ?obs sosa:hasFeatureOfInterest ?s .
}
```

Figure 3-20: SPARQL query for observations and association rules (Petrova et al., 2019)

An example resulting graph in Fig. 3-21 shows associationRule_1, which is linked to a sensor node (yellow node) and its two motif constituents. It is possible to retrieve SAX representations for each motif, as well as the observed variable and month they appear in. An appropriate user interface on top of those networks would allow the retrieval of discovered performance knowledge from the knowledge base.

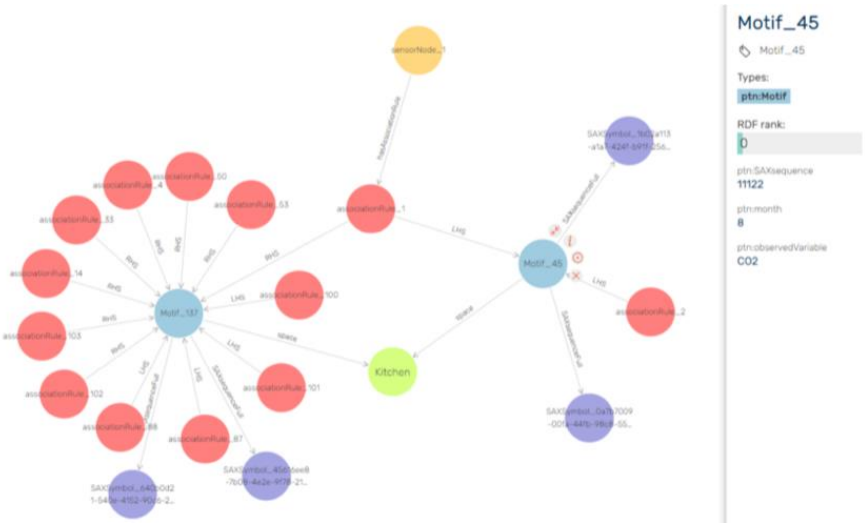

Figure 3-21: RDF graph with motifs, association rules and SAX representations (Petrova et al., 2019)

Yet, even though all performance patterns, rules and data are available in the knowledge base, there is no indication of what their meaning in terms of building performance is. Even though end users are able to retrieve any desired information in terms of performance, that would still only happen in the context of knowledge as a product of a data-driven discovery, as defined by Fayyad et al. (1996). To be fully useful, the system has to be able to provide an indication of what the discovered patterns and rules mean in terms of performance given the context they reside in (city, building, systems, occupants, etc.). Thus, the following Chapter 4 investigates how and to what extent expert interpretation can be added to the discovered motifs and rules in the knowledge base for context-aware evidence-based decision support.

For further details, please refer to Appendix A. Paper I, Appendix B. Paper II, Appendix C. Paper III and Appendix D. Paper IV: "Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data", "In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data", "Data mining and semantics for decision support in sustainable BIM-based design" and "From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches".

CHAPTER 4. KNOWLEDGE INTERPRETATION: CROWDSOURCING BUILDING PERFORMANCE PATTERNS

"All meanings, we know, depend on the key of interpretation."

George Eliot

As seen in both the literature review and the demonstrated building performance motif discovery, KDD approaches are highly capable of identifying patterns in large unfamiliar datasets. In fact, the accuracy with which machines perform such tasks often exceeds that of human domain experts. Yet, when it comes to analysis of the results, human domain experts reason based on high-level semantic abstractions to interpret them, while machines adhere to statistics and the applied models, which do not convey any explicit semantics. KDD algorithms can identify the frequent repetitive patterns, but cannot distinguish between meaningful and obvious patterns, or classify them in terms of usefulness. Domain knowledge and expertise, on the other hand, is a prerequisite for the interpretation and reasoning about the relationships between discovered patterns. Experts can also easily identify those patterns and/or rules that are most likely to be valuable and contain robust hidden knowledge. In accordance with the level of their expertise, domain experts are able to understand the meaning of the patterns in a given context. If these interpretations can be captured, they can create the backbone of a context-aware and semantics-aware decision support system.

In that relation, a possible solution in terms of design decision support could be to employ machine learning systems that are able to succinctly describe the discovered patterns in a way that a human domain expert can review and approve or reject them. Such a "tag team" approach would pair the extraordinary pattern recognition capability of machines with the domain knowledge of humans to add a level of intelligence that pushes the boundaries of conventional decision support to a human-centric system. Therefore, this chapter discusses the process of further enrichment of the semantic graph for contextualisation of the discovered building performance patterns and rules discovered and retrieved in Chapter 3, to support their interpretation by domain experts (Petrova et al., 2019b). The remainder of this chapter demonstrates the embedding of the domain expertise in the knowledge base through crowdsourcing and linked data techniques. Finally, the results, as well as the potential and challenges related to the presented approach are presented and evaluated.

4.1. CONTEXTUALISING AND FURTHER ENRICHMENT OF BUILDING PERFORMANCE PATTERNS

In terms of the domain knowledge needed for disambiguation of the discovered building performance insights, an important classification needs to be made. When referring to domain knowledge, it is essential to distinguish between domain knowledge in the sense of formal ontologies for explicit semantic demarcation of data, and domain knowledge in terms of human expertise needed to provide an indication of building performance. In this research, both concepts are actively used in several ways. First, domain ontologies are used for knowledge representation and storing of information in the semantic graphs constituting the knowledge base. On the other hand, expert knowledge is used for evaluation and disambiguation of the discovered performance patterns and rules. Finally, that expert knowledge as a concept has to be mapped to a formal ontology to be able to reside in the semantic graph and enable the targeted evidence-based process.

To further enrich the performance-enriched knowledge graph with a qualitative expert assessment containing possible meaning of the motifs and rules, some context need to be provided to make the interpretation both possible and accurate (Petrova et al., 2019b). Indeed, the performance patterns discovered in operational building data are frequently appearing regularities in known observed variables, but their appearance is caused by the influence of several factors. Those include, for instance, changes in external conditions, occupant behaviour, system performance, etc. Therefore, to provide such context and allow as accurate interpretation as possible, additional data is added to the performance-enriched semantic graph of Home2020. That includes weather data corresponding to the same time period as the collected data and for the precise geographic location (linking to OpenWeatherMap[18]); occupant data; consumption data related to the use of the heating system, the domestic hot water, and the use of appliances; HVAC system data, and HVAC design strategy for the building in accordance with the design brief requirements (Petrova et al., 2019b).

An overview diagram of the context-enriched version of the knowledge graph previously presented in Fig. 3-19 can be found in Figure 4-1. The context-enriched knowledge graph serves to further enrich the original data source and put the original data in an even broader context. This can be extended continuously as preferred and needed. Nevertheless, the more contextual data is presented to a domain expert, the more informative the performance patterns can be for this person, and the more useful this person's feedback and interpretation will become.

[18] [18] https://openweathermap.org

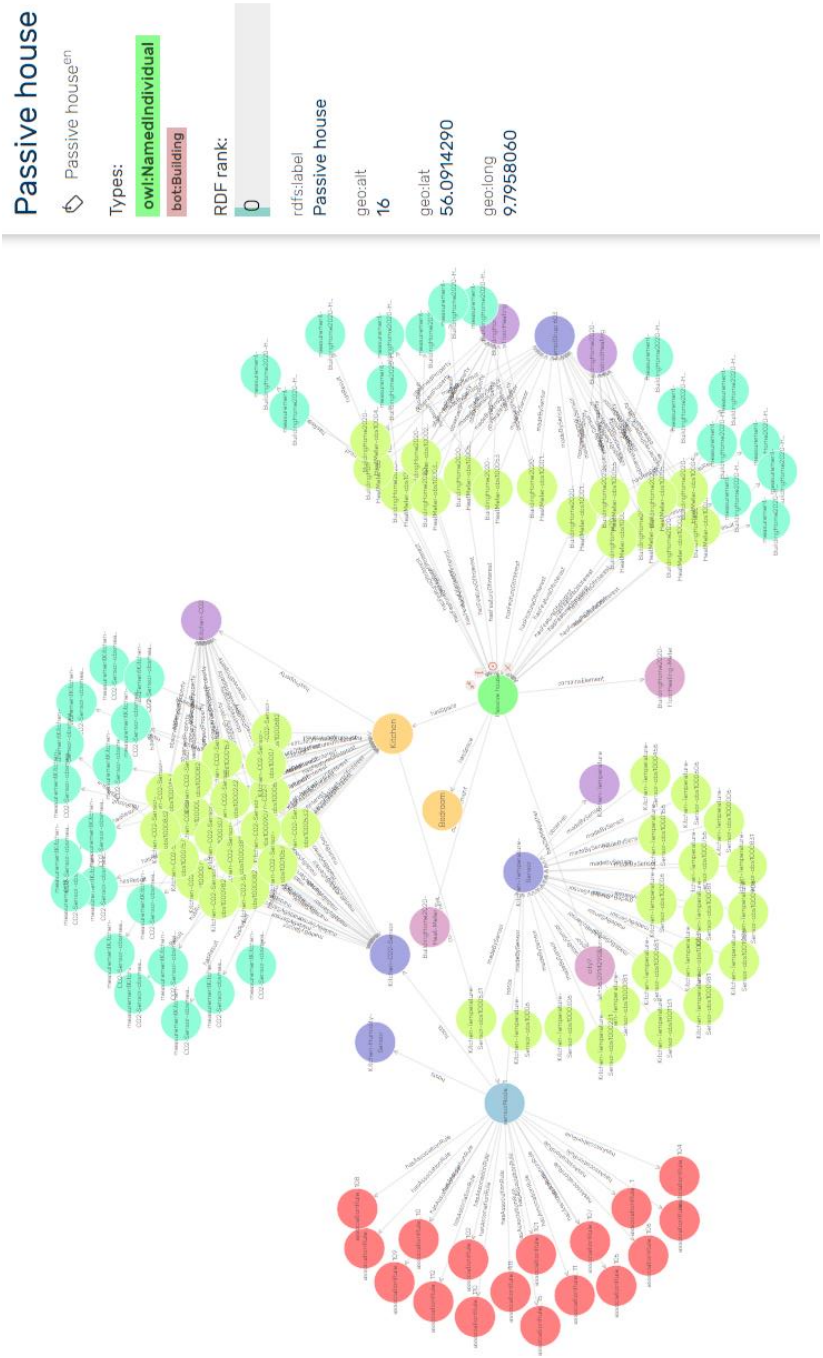

Figure 4-1: Context-enriched semantic building graph for expert interpretation of motifs and association rules (Petrova et al., 2019b)

Besides being used in its raw form to provide additional context to the discovered patterns, the newly added data can also be mined to obtain more and other kinds of insights, e.g. window opening and closing behaviour, energy consumption patterns, occupant profiles, anomalies in HVAC system operation, etc., which can also be added to the graph, to create an interlinked network of patterns and behaviours. Apart from data mining, other techniques can also be employed, e.g. annotation, NLP on input documents, feature recognition, and so forth. It is proposed here to use the same approach as recommended in Chapter 3, i.e. to include the original data in its original formats and include pattern recognition results in the semantic graph. All data is once again integrated through the same thin semantic integration layer for external access.

To be able to be interpreted and disambiguated, the discovered knowledge needs to be presented to domain experts in a structured way, that allows expertise to be continuously captured, updated and reused. A GUI therefore needs to be devised, which allows an expert to have a full tailored view of building performance patterns and their context, which is kept out of scope for this thesis. The following sections give an indication of the intended interaction between knowledge base and experts, so that they can provide their input and interpretations,

Important to note here is that this chapter aims to investigate the feasibility of capturing human domain expertise for disambiguation of knowledge discovery results and representing it in a semantically explicit way. The main focus is to define the most suitable approach and structure that fits the framework and knowledge base infrastructure created so far. The actual interpretation and an in-depth analysis of the building performance is out of scope.

The interpretations' credibility is of utmost importance and since expertise is a highly subjective matter, crowdsourcing techniques are considered as an alternative to rigid singular semantic annotation, to enable a level of statistical significance that testifies to an evidential character.

4.2. EMBEDDING DOMAIN EXPERTISE THROUGH CROWDSOURCING TECHNIQUES

4.2.1. CROWDSOURCING MECHANISMS AND PLATFORMS

The Semantic Web was conceived as a network that would allow machines to comprehend and respond to requests made by human users or other machines, as long as the data in that network is encoded with semantics (Berners-Lee, 2001). Naturally, the semantic richness of the data in that scenario is a key component. Yet, despite the presence of semantically rich data allowing to define objective knowledge (e.g. geolocations, product data, etc.), machines have significant limitations when the data is highly contextual, subjective and related to processes that are intrinsically performed better by humans (Xin et al., 2018; Acosta, 2014; Acosta et al., 2013). Such

subjective instances require semantic contextualization, disambiguation, interpretation, similarity matching, etc. The annotation of data is an essential aspect of knowledge interpretation and the richness of the semantic networks in Knowledge Base Construction (KBC) (Xin et al., 2018).

However, as stated in Petrova et al. (2019b), both conventional methods of human annotation and semantic web technologies in general are based on the antique ideal of a single correct truth, which does not respond well to the need of statistical significance and objectivity when it comes to data annotation. The concept of "crowd truth", on the other hand, aims to counteract the fact that human interpretation is subjective by postulating that collecting annotations of the same objects of interpretation across a crowd will reduce subjectivity, provide much more meaningful representations and reasonable interpretations (Aroyo, 2014). In other words, subjective knowledge has no documented ground truth but relies on dominant human opinion, which can be solicited from the (expert) crowd (Xin et al., 2018).

Howe (2006) coined the term crowdsourcing and defined it as *"the act of a company or institution taking a function once performed by a designated agent (usually an employee) and outsourcing it to an undefined and generally large network of people in the form of an open call".* According to Chiu et al. (2012), it originates in research on open innovation and co-creation, and allows to access intelligence and knowledge that are otherwise dispersed among many users (Schenk & Guittard, 2011). In that relation, Surowiecki (2005) states that the collective intelligence of the crowd, if the contributors refrain from communicating with each other, will converge on a more accurate solution to a problem than any of the expert members individually.

As a result, crowdsourcing has received major attention in the last decade in various domains. Research has investigated the use of crowdsourcing techniques for support of image recognition, product fabrication, rating systems, web development, etc. (Xiang et al., 2018; Petrova et al., 2019b). One of the most notable applications of such technologies is in design practices, including such based on AI, where crowdsourcing integrates human creativity with the machines' computational ability to produce designs (Xiang et al., 2018). In the Semantic Web domain, crowdsourcing has been applied as a means to obtaining high quality semantically annotated content, both in closed and open world settings. It has also proven to be a viable way of obtaining a sufficient number of human evaluators for qualitative evaluation tasks (Sack, 2014). Related research in the context of the Semantic Web also points to the use of crowdsourcing techniques for ontology engineering and knowledge base curation, validation and enhancement of knowledge and quality assurance of linked data (Sarasua et al., 2015).

The AEC domain has recently also begun to investigate the potential of such approaches in various contexts. Efforts include the use of crowdsourcing techniques for expansion of BIM-based construction material libraries through annotation of site

photo logs (Han & Golparvar-Fard, 2017) and creating annotations of construction workers based on building site video streams (Liu & Golparvar-Fard, 2015). In the infrastructure domain, crowdsourcing has been used for co-constructing and updating as-built BIM models, retrieving infrastructure operation and condition data, co-creating infrastructure sustainability and resilience, as well as infrastructure maintenance and rehabilitation (Consoli et al., 2015).

From a technical perspective, Blohm et al., (2018) state that crowdsourcing platforms can be distinguished according to several criteria. The main differentiation is based on the diversity of the contributions and the ways in which these are aggregated. In terms of diversity of contributions, crowdsourcing platforms can be divided into homogeneous (crowd contributions are characteristically identical) and heterogeneous (crowd contributions differ in nature and quality). As of aggregation, research distinguishes between selective contributions (value is derived from individual contributions) and integrative ones (value is derived from the entirety of all contributions (Blohm et al., 2018).

Crowdsourcing in the context of this research project clearly points at the need of characteristically identical contributions derived from the entirety of all contributions (homogeneous and integrative) (Petrova et al., 2019b). Blohm et al. (2018) define this crowdsourcing type as "Information Pooling", which is based on additive aggregation of distributed information and aims to integrate diverse opinions, assessments, predictions or other information from contributors. It is also important to underline the significance of the expert factor. Interpretation of building performance patterns requires specific high-level expertise, and the use of crowdsourcing is intended in that context. Yet, crowdsourcing to expert crowd/ end users here implies indoor environmental quality and building performance professionals, who are also familiar with the design process and/or are a part of it as end users of the envisioned decision support system (Petrova et al., 2019b).

Therefore, the remainder of this chapter discusses the implementation of crowdsourcing techniques for interpretation of building performance patterns by an expert crowd.

4.2.2. CROWDSOURCING BUILDING PERFORMANCE PATTERNS

This section gives of overview of the proposed crowdsourcing platform for retrieval of building performance and indoor environmental quality domain expertise for disambiguation of patterns discovered in operational building data. This section hereby relies on the dataset that was already presented in Chapter 3 for the Home 2020 case and further enriched with contextual data as presented in Fig. 4-1.

First of all, retrieval of domain expertise requires an environment in which a human domain expert can work and assess building performance patterns. As a result, a GUI is needed in which the contextualised data needs to be presented to an expert end user.

It is key here that the proposed crowdsourcing tool aims at annotating the association rules discovered in building performance data and not all the other data, which already has much of the desired semantic demarcations and classifications. Thus, the semantic enrichment of association rules is the key objective of the crowdsourcing task in this work. The development of the GUI itself is out of scope in this work, but it is important to note that special attention has to be paid when designing a GUI for selection of predefined semantic categories, because the functionality of the system will have a direct effect on the quality of the crowd contributions. This research focuses on defining the necessary underlying infrastructure that enables the crowdsourcing effort.

What requires semantic annotation?

In the case of the Home2020, several hundred association rules have been retrieved in sensor data from three specific months (January, April, August), three rooms (bedroom, living room, kitchen) and three observed indoor environmental quality variables (Temperature, CO_2, Relative Humidity) (see also Chapter 3). Figure 4-2 shows five of the hundreds association rules including their measures of "interestingness": support and confidence (Petrova et al., 2019b). Each rule contains the IDs of the motifs that constitute the rule and the numerical values for support and confidence of the rule. The level of support hereby equals the number of transactions that contains both the antecedent and consequent of the rule. The confidence of a rule is an expression of how often that rule is found to be true, which is calculated by the number of transactions that contain the antecedent and consequent of the rule, divided by the number of transactions that contain the antecedent (Petrova et al., 2019).

```
452 ==> 489 #SUP: 1 #CONF: 1.0
453 ==> 485 #SUP: 3 #CONF: 0.6
454 ==> 481 #SUP: 1 #CONF: 0.5
456 ==> 484 #SUP: 2 #CONF: 0.6666666666666666
457 ==> 488 #SUP: 1 #CONF: 1.0
```

Figure 4-2: An excerpt of association rules found in data from Home2020 (Petrova et al., 2019b)

For example, the rule 453 ==> 485 in Fig. 4-2 means that whenever pattern 453 is found, pattern 485 is typically also found. In the available data, motifs 453 and 485 co-occur 3 times (support = 3), and, since the antecedent (pattern 453) appears 5 times in total in the analysed dataset, confidence is equal to 3 divided by 5, and thus 0.6. In other words, three out of five times (60%), pattern 453 co-occurred with pattern 485; the other two times, pattern 453 co-occurred with a different pattern (Petrova et al., 2019b).

The precise character of the same example association rule (453 ==> 485) is visualised in Fig. 4-3 based on the SAX representations of the motifs. Patterns 453 and 485 represent two different SAX strings, namely 55544 (Relative Humidity) and

5555544444 (Temperature). The symbols in the SAX strings hereby belong to the specific intervals found earlier in the SAX computation step for each observed variable (see Chapter 3). For Relative Humidity, the SAX symbol '4' represents the interval [39.05,41.61] and '5' represents the interval [41.61,44.39] with a unit of measure [%]. For temperature, the SAX symbol '4' represents the interval [24.73,25.35] and '5' represents the interval [25.35,26.03] with a unit of measure [°C]. In other words, that particular association rule means that whenever the indicated interval sequence in Relative Humidity occurs, there is a 60% chance that the corresponding interval sequence in Temperature occurs (Petrova et al., 2019b).

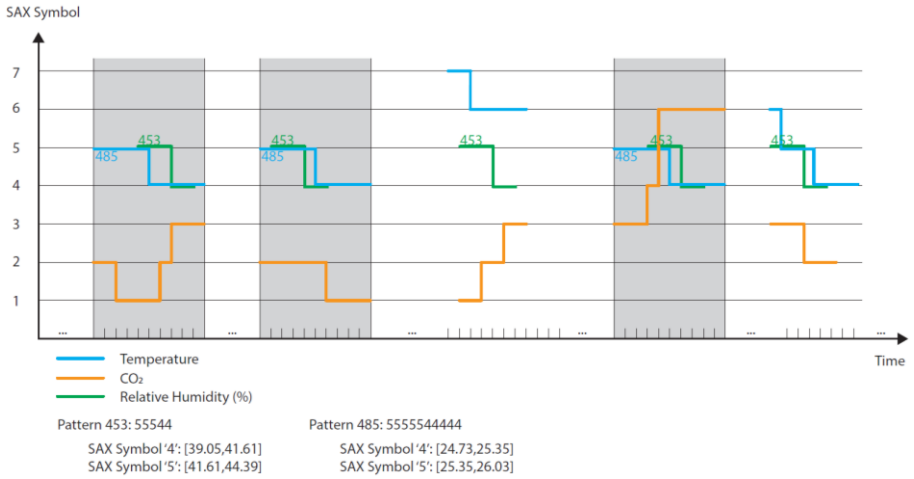

Figure 4-3: A visualization of an association rule in indoor environmental quality data, the motifs that it consists of and their corresponding SAX representations (Petrova et al., 2019b)

The next sections indicate how input from domain experts can be retrieved and included in the knowledge graph to interpret the meaning of association rules such as the one described above. As stated in Petrova et al. (2019b), a two-fold methodology is hereby applied, which targets semantic annotation of building performance patterns in a first round and relies on crowdsourcing techniques that utilise those annotations to evaluate the building performance patterns and transform them into a valuable decision support mechanism. Both techniques are thereby employed together as part of the same crowdsourcing system (Petrova et al., 2019b).

Semantic annotation of building performance patterns: principles and ontologies

Upon the presentation of an association rule to a domain expert, this person can identify certain features about the association rule and annotate them directly, as part of the semantic graph. In such case, original data, discovered motifs, and expert interpretation by annotation are all stored in the same graph, together with the additional contextual information or links to external information (Petrova et al.,

2019b). This research effort aims for both semantic classifications and human-readable descriptions to provide a more unambiguous and informative interpretation, aiming to capture what would come closest to performance "stories" (Heylighen et al., 2007). The reason for that is to avoid "single truth" annotations (semantic classification only) and be able to capture as much as possible form the tacit concept and the context in the description following the annotation. Subsequently, the use of semantic relations between descriptions may also be considered as an alternative retrieval approach, similar to the approach defined by de Vries et al. (2005). Furthermore, the use of a broader range of information retrieval resources, such as keyword searches, implementation of auto-suggestion services for suggestion of potentially suitable semantic annotation entities that fit the user context and input best, etc. is also possible.

As stated by Petrova et al. (2019b), the annotations by the experts are collected through the crowdsourcing platform and stored directly in the knowledge graph for reference. The goal here is once again to rely on available and proven ontologies for such expert-defined annotations. Of course, a lot of contextual information is already available about the discovered association rules and motifs. As with any semantic annotation system, human annotations by an expert will lead either to the addition of classifications, and/or to the addition of "stories" and more descriptive comments. While the former is much more reusable by a machine, especially in the information retrieval steps, the latter is much more informative, in the sense that such description tags include a more elaborate interpretation of each association rule. Such description tag can however only be fully utilised by a human end user (human-centred). Tags reflect the experts' personal interpretations of the world and are therefore not normalised for machines (Petrova et al., 2019b).

A number of options is available for storing the classifications and descriptions. One option is the use of the Review ontology[19]. This ontology allows to use classes such as Comment, Review, Feedback, etc. Key in this approach is that the ontology allows to link a Review directly to a "work". This Review is then central for adding more details, such as comments and feedback on that review. Agents or people are hereby modelled using the FOAF ontology[20] (Petrova et al., 2019b). The above suggested tagging approach (classifications and descriptions) could rely on the Review ontology. Alternatively, it is possible to rely on the Review and Commenting mechanisms provided by the schema.org ontologies[21]. In this case, Reviews and Comments can be directly linked to the schema:CreativeWork class. Instead of using the FOAF ontology for defining people, the schema:Person class can be used. Furthermore, the ontology provides the option to store votes (e.g. schema:upvoteCount), and is more flexible, in

[19] http://vocab.org/review/

[20] http://xmlns.com/foaf/0.1/

[21] https://schema.org/Review

the sense that Reviews, Comments, and CreativeWorks can be combined in a several ways, with the possibility of adding metadata to each (agent, about, dateCreated, text, etc.) (Petrova et al., 2019b).

When applying the schema.org inspired approach to the targeted semantic tagging / annotation system, the data model for the annotation of an association rules resembles the diagram presented in Fig. 4-4 (Petrova et al., 2019b).

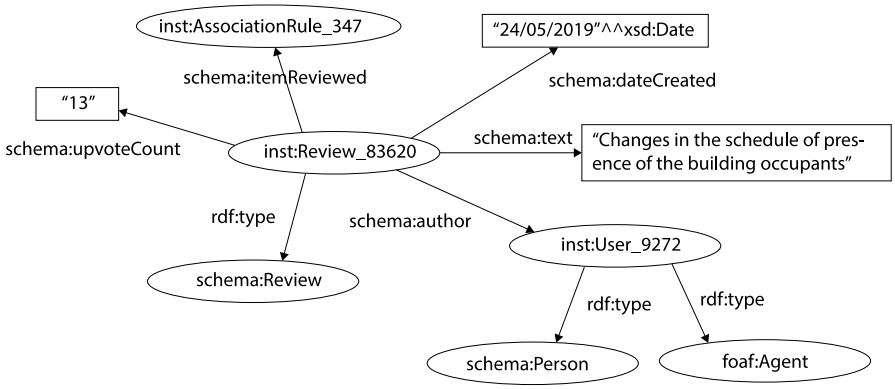

Figure 4-4: Data model for semantic annotation and interpretation of association rules discovered in operational building data (Petrova et al., 2019b)

It has to be mentioned that this approach so far results only in the addition of reviews with expert-defined text descriptions as input. This approach is useful, but alone it does not provide semantically definitive tags or classifications, which can be used in information retrieval. Therefore, it is proposed here to extend the above work with the possibility to add semantically defined tags (classification) (Petrova et al., 2019b).

Semantic annotation tags

When implementing a tagging system, five main categories can be used to group or classify tags that reflect the most usual causes of any pattern (motif or discord) appearing in sensor observations from buildings in operation. Those are typically related to dynamic parameters that have an observable direct effect on the behaviour of a building, which are hereby used as main classification tags, namely (1) external conditions, (2) occupant behaviour, (3) system performance, (4) design and (5) construction. When tagging and classifying association rules, any comment resides under one of these five main tags (Fig. 4-5) (Petrova et al., 2019b).

Under each of these classification tags, a number of standard tags are available, which can be selected by the domain expert for annotation of an association rule. Furthermore, the system allows to add new, previously undefined tags, as deemed necessary by the domain expert (Petrova et al., 2019b). Over time, the number of

default available tags can be revised in order to better respond to the tagging behaviour.

As all tags need to be collected, it is suggested to store all tags into a separate graph, to which additional tags can be added as preferred. Ideally, a user does not need to devise new tags continuously, but instead can rely on the tags available in the defined AllTags graph or a Tag Dictionary. As a result, a number of tags are available under each of the given categories (see subClassOf tree structure in Fig. 4-5), which can be selected to complete the building performance pattern reviews (Petrova et al., 2019b).

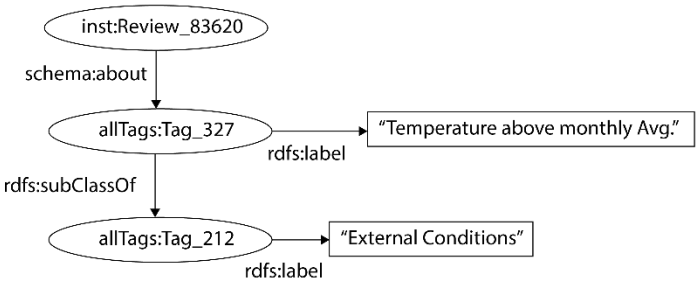

Figure 4-5: Semantic tags for classification of expert reviews of association rules (Petrova et al., 2019b)

Crowdsourcing Platform-User interaction

The above sections document the data model that can be used for semantic annotations and tagging by domain experts. Of course, this data model needs to be embedded in a web-based application that allows to present domain experts with association rules and enables them to provide input about association rules stored according to that data model (Petrova et al., 2019b).

Figure 4-6 presents an interaction diagram that indicates how feedback and comments are retrieved from the user's perspective. As shown in the diagram, ARM nodes are retrieved from the knowledge base, each of them identified by a URI. That includes retrieval of the relevant contextual information available in the graph (Steps 1-3). In case one or more reviews are already available, those are presented to the user as well. This provides the option for the domain expert to add upvotes to the already available reviews, depending on whether or not the reviews are considered reasonable and based on the provided rule attributes and context (Step 4a). At any time, a domain expert is able to assign a new review to the association rule, to which metadata is attached (user metadata, date, profile, etc.) (Step 4b). For each review, a description is added, as well as a semantic tag from the repository of tags (Steps 5a, 5b). All reviews and comments are stored in a separate graph, yet linked to the particular association rule's URIs and the user profile URIs, as indicated in the data model outlined in the previous section (Petrova et al., 2019b).

For each tag that is added to an association rule by a user, for whom user details are available after login, a new tag is added, including the associated user profile, a date, and a human-readable description, as can be seen in Fig. 4-6. In other words, motif and ARM nodes in the knowledge base are retrieved and obtain additional metadata, including classification, user metadata, profile, etc. That can be put in a separate graph connected to the same URIs, including human evaluation (Petrova et al., 2019b).

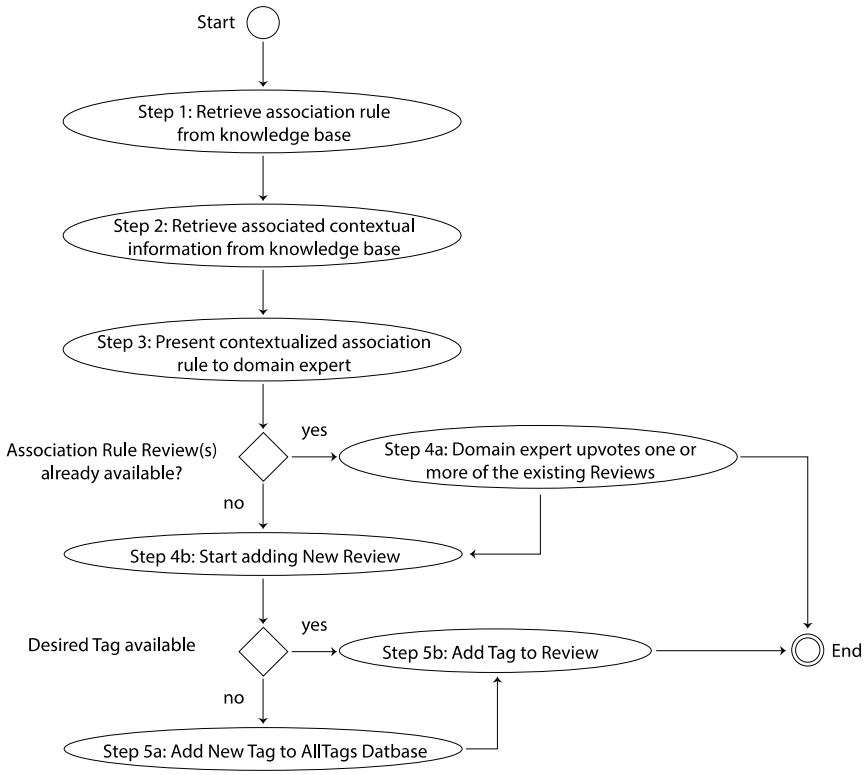

Figure 4-6: An interaction diagram showing the steps that a user undertakes in retrieval, reviewing, annotation and tagging of association rules (Petrova et al., 2019b)

4.2.3. FROM DIRECT BELIEF TO KNOWLEDGE

Figure 4-7 presents the overall crowdsourcing setup and the way the outlined system fits into the overall research framework presented throughout the thesis and the system architecture outlined in Chapter 3.

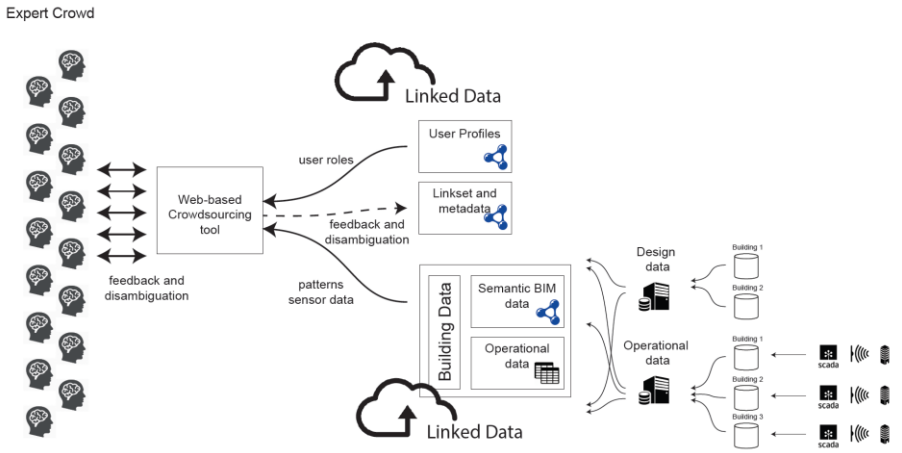

Figure 4-7: The proposed crowdsourcing system in the context of the overall research framework (Petrova et al., 2019b)

In principle, the crowdsourcing system for semantic annotation and interpretation functions as follows. The knowledge base hosts all association rules discovered in the data. Each of them can be visualized, including the related context in which it resides. When an expert logs in, and activates their user profile, they can browse the available association rules. The expert then has the option to express their belief by either defining a new meaning (annotation) of a rule and classify it with semantic tags, or upvote and refine existing interpretations (review), which get stored in the graph. All this data is stored as part of the graph, including a reference to the URI of the domain expert user. Eventually, under the impact of the crowd of domain experts, the most interesting patterns become clearly visible, ideally also including comments and annotations that can be useful for information retrieval from any future design environment to clarify the impact of particular design decisions or causalities in building performance.

Naturally, the above-defined semantic annotation and tagging mechanism is only as good as the provided input- the classification tags and the expert interpretations (descriptions). Even with semantics attached to the association rules, for an end user or for a machine, it is still very difficult to find out which patterns are of higher value. A solution to that could be a semantic enrichment system that focuses less on semantic annotations and interpretation and more on annotating association rules directly, primarily based on interestingness (Petrova et al., 2019b). However, the known measures of interestingness (i.e. support, confidence, lift, etc.) are also only a partial and subjective factor decided by the analyst. Instead of only adding specific semantic annotations, it might be useful to let domain experts log in, and browse association rules, without being pointed to rules classified as interesting only based on the associated support and confidence values. Considering the nature and value of human expertise, it might suffice to visually indicate where co-occurring motifs (or

association rules) happen in context, and experts might be able to indicate precisely which the interesting co-occurrences and rules are. Providing such expert input by adding upvotes directly to the association rule may be sufficient. (Petrova et al., 2019b).

With this addition of a direct upvoting mechanism for association rules, Fig. 4-8 hereby showcases the full crowdsourcing principle proposed here. The functioning of the mechanism can be summarised as direct crowdsourcing with three main expert contributions (Petrova et al., 2019b):

(a) **Input:** Domain expert users (User 1 and User 2 in Fig. 4-8) provide their beliefs and input about new rules or refine and update already existing knowledge. The users choose freely which entities to operate on without predefined suggestions or other constraints. The input is stored in the knowledge base.
(b) **Review:** Other experts from the crowd (Users 3-10 in Fig. 4-8) also provide their input in the form of new annotations and tags, or interact with the existing ones, thereby upvoting or refining the existing interpretations. The input is also stored and analysed against the existing knowledge base. The experts receive feedback about any internal technical inconsistencies inside their update in a real-time manner.
(c) **Upvote:** The experts upvote triples suggested by other experts. Users upvote annotations in case their belief confirms an existing annotation from another expert.

Important to note here is that the refinement of an existing annotation does not imply override of existing annotations. Currently the implementation limits to the ability to upvote and review by adding a new description. In that sense, experts cannot override or update annotations provided by other members of the crowd, only upvote them. Another option would be to compute updates dynamically based on level of compliance with the existing knowledge base. Once a certain threshold of a number of upvotes, equal or higher than the existing ones has been reached, the new interpretation is automatically integrated into the existing knowledge graph. However, such an implementation is out of scope for this work.

As concluded in Petrova et al. (2019b), the proposed crowdsourcing approach for interpretation and annotation of association rules can be beneficial because it allows the expert crowd as users to work directly with the existing hierarchy of classes and no other entities. In addition, the domain experts do not need any information or familiarity with the existing knowledge base to be able to provide new input. The data necessary for contextualisation of the rules is retrieved along with the retrieval of the association rule (Step 1-3 in Fig. 4-6). That makes the approach suitable for large knowledge bases. Furthermore, Semantic Web technologies and reasoning mechanisms can be of utmost value for analysing the experts' input, govern quality

and validity and help avoid contradictions between annotations. Such a system may have an implicit function aiming to use the provided interpretations, provide feedback and serve as an educational mechanism for the crowd (Petrova et al., 2019b).

Finally, of importance is also the aggregation of interpretations and tags over time. Essentially, the layer of the knowledge base consisting of the semantic annotations has to accumulate to a point where it becomes statistically significant and useful in terms of decision support. At a later stage, the potential of self-learning systems and reasoning agents can be explored further in terms of self-annotation of expert interpretations. In any case, the discussed functionalities have to be tested with an actual crowd of domain experts for feasibility and usefulness (Petrova et al., 2019b).

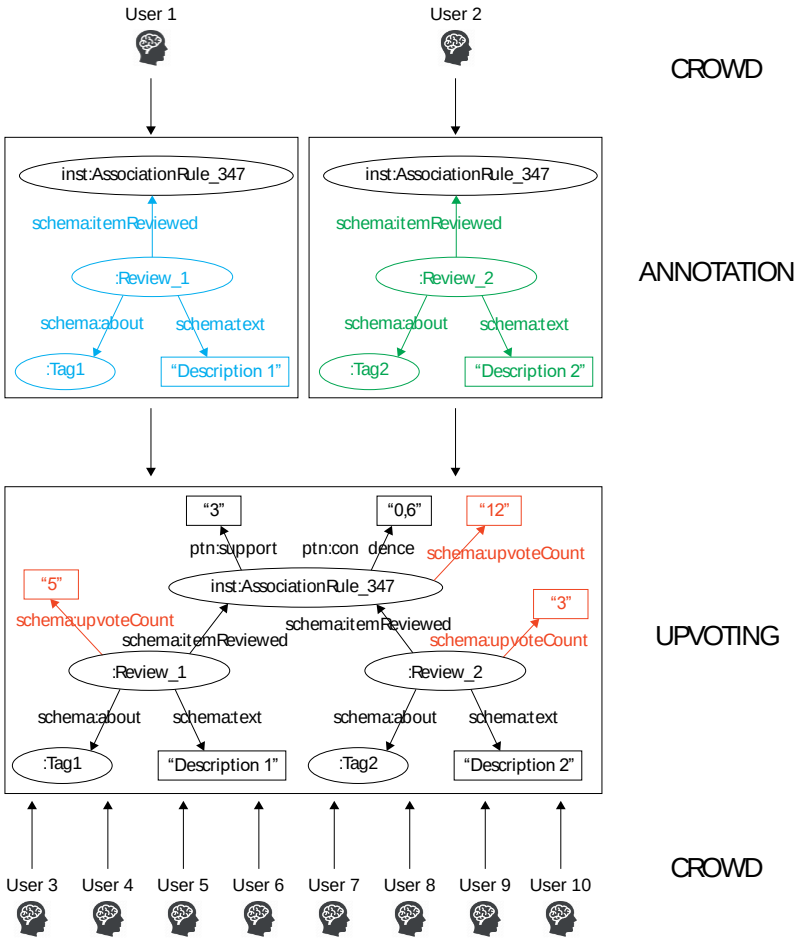

Figure 4-8: A snippet from the graph containing the expert crowd annotations and reviews of discovered association rules and the crowdsourcing process (Petrova et al., 2019b)

4.3. CHALLENGES AND LIMITATIONS

The initial evaluations of the proposed system show that using crowdsourcing techniques for disambiguation and interpretation of knowledge discovered in operational building data holds significant potential. That is particularly valid in terms of removing the long-standing boundary between the output of traditional machine learning approaches for knowledge discovery and the ability to reuse those results in a way that is meaningful to both humans and machines.

However, certain challenges need to be considered and further addressed. First of all, even though the interpretation of the discovered knowledge is embedded in the knowledge base, that by itself does not synthesize solutions in terms of design decision support and these have to still be carefully devised by another system tailored towards recommendations for users, that is built on top of the enriched knowledge base(s) and the design team itself (Petrova et al., 2019b).

Furthermore, even though the crowdsourcing system in the context of this research is explicitly designed to rely on expert crowds and not layman users, it still has to be assumed that the quality of the contributions may vary. That means that an additional verification and validation layer also has to be considered, which may require additional rounds of expert reviews or a rule-based system (Petrova et al., 2019). The actual usefulness of the contributions also has to be assessed. Over-engagement in that sense should be considered as a potential issue leading to the crowdsourcing process being "too successful", i.e. too many crowd contributions, few of which having real value and being worth implementing in decision support. The valuable output of the interactive knowledge sharing may, in that case, be lost or become harder to identify amongst all contributions. The opposite challenge is, of course, also possible: too little engagement and too few contributions available to be able to provide any substantial basis for knowledge retrieval.

All these challenges need to be addressed, so that the value of the crowdsourcing effort based on Semantic Web technologies can be harvested in performance-oriented design practice. In that relation, one of the most important elements of the envisioned evidence-based and user-centred design decision support is using the created knowledge base in a way that allows reaching the design team and make an impact. Therefore, the next chapter presents the final effort in this research, namely bringing back the discovered and interpreted knowledge to the design team in the form of user centred dedicated recommendations.

For further details, please refer to Appendix E. Paper V: "Crowdsourcing building performance patterns for evidence-based decision support in sustainable building design".

CHAPTER 5. CLOSING THE LOOP BETWEEN BUILDING OPERATION AND DESIGN WITH KNOWLEDGE-BASED DECISION SUPPORT

"Knowledge is of no value, unless you put it into practice."

Anton Chekhov

The previous chapters presented the full transformation of operational building data and archival project data into a rich knowledge base capable of providing evidence-based decision support to a design team in a performance-oriented design process. That transformation includes the definition of the main types of building data, the analytical approaches that can be used to extract valuable insights and the ways in which semantic web and linked data technologies can be used for the representation and retrieval of those insights. As the explicit meaning of the discovered patterns typically remains unknown to the machine, the thesis also approached the challenge of contextualisation, disambiguation and interpretation of the knowledge discovery results, thereby enabling the semantic integration of the discovered building performance insights into rich knowledge bases, able to serve as an underlying basis for design decision support.

As previously stated, the power of AI technologies lies in the ability to enhance human decision-making. It was also argued that the richness, structure, and accessibility of the knowledge bases are essential to the decision-making processes and the related systems. In that relation, the final part of this research effort aims for bringing valuable knowledge from the knowledge ecosystem back to the design team through meaningful recommendations based on various levels of similarity with the current design context of the team.

5.1. LINKED DATA-BASED RECOMMENDER SYSTEM FOR IMPROVING SUSTAINABLE DESIGN DECISION-MAKING

The main hypothesis of this final part is that high-quality recommendations require that (1) the profile and context of the design team is appropriately analysed, (2) the most relevant cases from the knowledge bases are found, and (3) the retrieved knowledge is effectively communicated to the design team. Therefore, the following sections of this chapter analyse how the created infrastructure and knowledge base can be put into use, how the design team member user profiles can be built and benefit

from the system, how the user feedback is handled and how linked data-based recommendations can be generated. The semantic data modelling approach follows the same practices as outlined in the previous chapters. Figure 5-1 lists all namespaces and prefixes further used throughout this chapter.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix bot: <https://w3id.org/bot#> .
@prefix buildings: <https://www.example.com/data/buildings/> .
@prefix people: <https://www.example.com/data/people#> .
@prefix ls: <https://www.example.com/voc/linkset#> .
@prefix bmeta: <https://www.example.com/voc/buildingmetadata#>
    .
```

Figure 5-1: Namespaces and prefixes used throughout the chapter (Petrova et al., 2019a)

Following the above described principles results in the conceptual system architecture in Fig. 5-2. The following sections explain the proposed architecture in more detail, thereby focusing on how design team user profiles can be built and benefit from the system, how the user feedback is handled and how the recommendations are generated and retrieved (Petrova et al., 2019a). Finally, the thesis demonstrates an initial implementation of a linked data-based recommender system by applying the concept of Linked Data Semantic Distances (LDSD) proposed by Passant (2010).

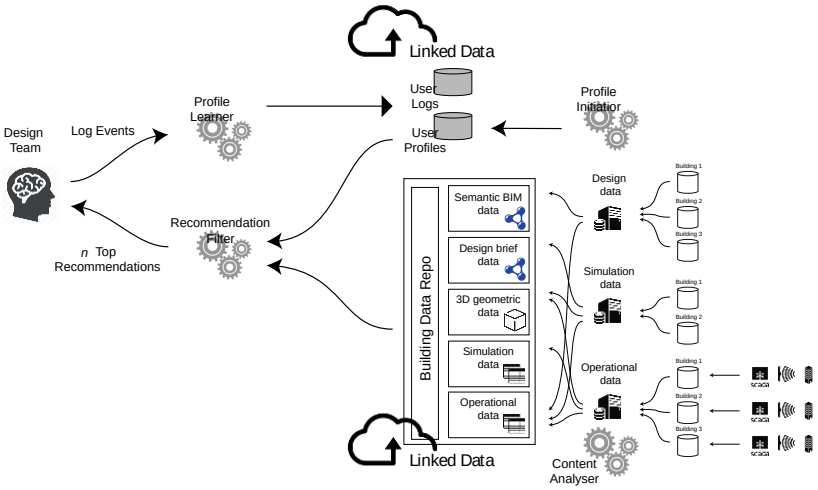

Figure 5-2: System architecture for a LOD-based recommender system in performance-oriented building design relying on knowledge bases (Petrova et al., 2019a)

5.1.1. USER PROFILING AND FEEDBACK

The first fundamental concept in the defined system architecture is establishing the user (design team/professional) context and profile, as well as the feedback that the user provides when interacting with the system (Petrova et al., 2019a). All these features are essential to the performance of the recommender system. In terms of user profiling, the system is conceived according to a methodology similar to the one proposed by Boratto et al. (2017).

As outlined in Petrova et al. (2019a), a Profile Initiator component fills a dedicated RDF-based User Profile Store at user registration. Similarly to the crowdsourcing effort described in Chapter 4, all user profiles here are also RDF-based and modelled using the FOAF ontology, thereby identifying each user and their metadata (Fig. 5-3). In fact, the same User Profile Store may be used for the recommender system and the crowdsourcing tool.

```
people:EkaterinaPetrova
  a foaf:Person ;

  foaf:name "Ekaterina Petrova"^^xsd:string ;
  foaf:givenName "Ekaterina"^^xsd:string ;
  foaf:familyName "Petrova"^^xsd:string ;
  foaf:nick "epetrova"^^xsd:string .
```

Figure 5-3: Example of people profile data, modelled according to the FOAF ontology (Petrova et al., 2019)

As soon as the user starts using the system, they are served recommendations through the Recommendation Filter component. All actions that the user undertakes, in direct interaction with the recommender system (e.g. clicking a recommendation, loading a recommendation, viewing a recommendation, clicking a 'Like' button, etc.), are logged through a Profile Learner component (Fig. 5-2) (Petrova et al., 2019a). Such actions can be identified by tracking clicking behaviour, eye tracking, etc. A full investigation of the user interaction with the system, as well as development of the GUI of the system are out of scope in this work, however, the main infrastructure, functionality and resulting recommendations are further discussed.

The Profile Learner component feeds user profile data and user logs back into the back-end of the recommendation system, where the User Profile store and the User Log store are located (Fig. 5-2). Thus, the User Profile store gets modified incrementally, in response to the interactions by the end user, most influential of which would potentially be the used recommendations responding directly to specific design requirements and performance targets. The feedback from the user interactions goes into User Logs and User Profiles, but the link between specific user profiles and relevant items in the Building Data Store are kept (binding linkset), thereby aiming to enable context-aware recommendations (Petrova et al., 2019a). In other words, as

further mentioned in Petrova et al. (2019), links between user profiles and building identifiers are kept in a separate RDF linkset (Fig. 5-4), which serves as a hash table with identifiers from the User Profile store and the building data repository. In this particular example, only the ls:likes relation is showcased, but multiple other relations may also be specified based on the ways in which user interaction and feedback are tracked.

```
people:EkaterinaPetrova
  ls:likes buildings:building_987d706d-877a-4b1d-80f6-6
    ee89d856319 ;
  ls:likes buildings:building_af41d889-f50c-456e
    -9625-96655150838d .
```

Figure 5-4: Linkset between buildings and people based on the ls:likes relation (Petrova et al., 2019a)

The same building principle is also applied to the Building Data Store, User Profile Store, and Linkset Store when adding implicit data about buildings in the building data repository. The buildings can be enriched with metadata tags such as buildingType, designedBy, energyLabel, sustainabilityCertificate, etc. to form categories of design references, to compose queries in the database, to sort search results according to specific criteria, etc. (Petrova et al., 2019a). The example in Fig. 5-5 only uses several simple metadata tags, but multiple other metadata tags related to, for instance, geographic location, building occupancy, mined performance patterns, energy source, etc. can also be added. Of course, in this work, the metadata about the buildings and patterns retrieved by crowdsourcing (see Chapter 4) is considered to be extremely relevant data, which can be used by the recommender system in addition to the more simple tags which are used here (Fig 5-2) for explaining the potential for initial semantic enrichment.

```
buildings:building_00dd6c87-6a6e-f482-7490-e6613659708a
  a bot:Building ;
  bmeta:buildingType bmeta:theater ;
  bmeta:designedBy people:architectX ;
  bmeta:energyLabel bmeta:A ;
  bmeta:sustainabilityCertificate bmeta:LEEDPlatinum .

buildings:building_2e0dcc1c-b981-4c47-adb4-2b9887f10481
  a bot:Building ;
  bmeta:buildingType bmeta:theater ;
  bmeta:designedBy people:architectY ;
  bmeta:energyLabel bmeta:A ;
  bmeta:sustainabilityCertificate bmeta:DGNBGold .
```

Figure 5-5: Building data enriched with metadata tags (Petrova et al., 2019a)

To summarise, the system holds four RDF-based data stores, i.e. User Log Store, User Profile Store, the Building Data Store, and the Linkset Store (Petrova et al., 2019).

The Linkset Store maintains all the linksets between data in the other three stores. Eventually, that combination of data stores and a Linkset Store allows to retrieve relevant information with user queries. For instance, user queries may target retrieval of buildings of particular types, category, energy label type, etc. As indicated in fig. 5-6, the metadata (bmeta) tags showed in Fig. 5-5 can be used. Yet, as further stated by Petrova et al. (2019a), user preferences (Linkset Store) or user profiles (User Profile Store) can also be included in the queries. To achieve that, the metadata (bmeta) tags are used (Fig. 5-6). As further stated in (Petrova et al. (2019) user preference (Linkset Store) or user profile (User Profile Store) data can also be included in the queries. Metadata retrieved through the crowdsourcing tool in Chapter 4 may also be used in this information retrieval step.

```
SELECT *
WHERE {
    ?b a bot:Building .
  ?b bmeta:buildingType bmeta:theater ;
    ?b bmeta:energyLabel bmeta:A .
}
```

Figure 5-6: SPARQL query for buildings of a particular building type with bmeta tags (Petrova et al., 2019a)

The linkset that binds building data with metadata, user data, etc. can in a next phase be used by the Recommendation Filter to further optimize the recommendations it provides to end users, i.e. the *n* top recommendations become more user-tailored depending on the user context (Fig. 5-5) (Petrova et al., 2019a). The following section will further discuss how these links are used by the Recommendation Filter.

5.1.2. GENERATING RECOMMENDATIONS

Instead of only relying on metadata tags and user queries that can be sent from the end-user environment, the recommender system should also be able to recommend buildings that are semantically close to a building that is considered to be most relevant to an end user at some point in time based on the information learned about the user profile and their context (Petrova et al., 2019). This requires a “push” system architecture (suggestions by the system based on user interaction), rather than a “pull” system architecture (questions by the user).

To provide recommendations, recommender systems in general rely on a certain level of similarity between concepts. In this case, the semantic backbone of the knowledge base means that the computation of recommendations can be based on the semantic relatedness between concepts, also defined as semantic distance. Figure 5-7 shows the principle of the generation of recommendations to end users based on their interaction with the system and context.

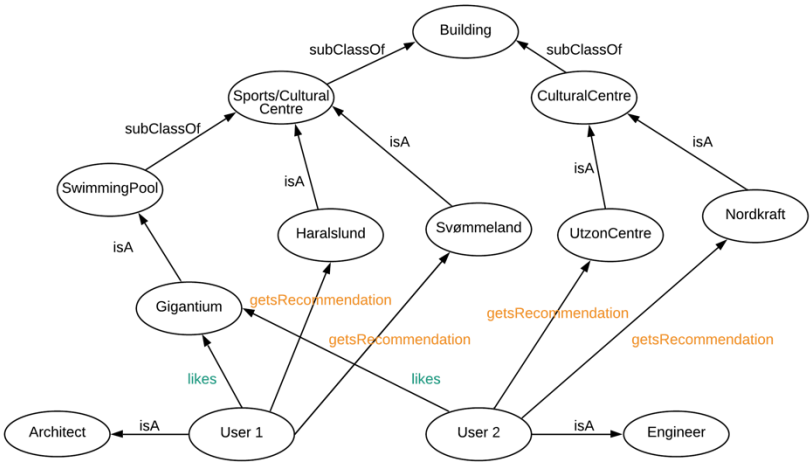

Figure 5-7: Graph-based recommendations to users based on interaction and context

Figure 5-7 also indicates recommendation activity based on semantic relatedness between concepts. Passant (2010) introduced a set of measures that can be used to determine the 'Linked Data Semantic Distance' (LDSD) between two concepts in the context of graph-based recommendations. The LDSD values range between 0 and 1, and the smallest distance implies the highest level of similarity between resources (semantic closeness). Passant (2010) hereby distinguishes between Direct, Indirect, and Combined Semantic Distance ($LDSD_d$, $LDSD_i$, $LDSD_c$ respectively), each either weighted or not. The distance $LDSD_d$ considers strictly the direct links between resources, both incoming and outcoming. Since one of the biggest values of linked data in general resides in the indirect links between concepts through connections and other concepts, the author also introduces indirect $LDSD_i$, which is based on indirect links between resources. Finally, $LDSD_c$ combines both (Passant, 2010). Recommender systems use these measures to find out what else users may like based on their profile, search behavior, favorites, likes, etc. The smaller the semantic distance between two related concepts, the higher the related concept is ranked in the set of *n* top related recommendations in the system outlined in Fig. 5-2 (Petrova et al., 2019a).

In principle, semantic distance can be computed using all of the outgoing and incoming links of two concepts, which are bmeta and bot links in the simplified case used in this chapter. In a fully contextualised graph enriched with domain expert input through crowdsourcing techniques, a lot more diverse links can be considered between two distinct concepts.

In the case considered here, for instance, different buildings might be attributed to be "theatre" buildings, which connects them to the same node for the bmeta:category predicate, and makes them semantically closer to each other (Petrova et al., 2019).

That concept is also present in Fig. 5-7. In this research effort, the method proposed by Passant is used to determine the indirect semantic distances $LDSD_i$ between buildings in the knowledge base tagged with the discussed bmeta tags. The semantic relatedness between buildings in the knowledge base is thereby determined with an implementation of the following $LDSD_i$ equation:

$$LDSD_i(r_a, r_b) = \frac{1}{1 + C_{io}(n, r_a, r_b) + C_{ii}(n, r_a, r_b)}$$

Figure 5-8: Indirect semantic distance $LDSD_i$ (Passant, 2010)

The above equation is based on the definition provided by Passant (2010), which states that *"C_{io} an C_{ii} are functions that compute the number of indirect and distinct links, both outcoming and incoming, between resources in a graph G. $C_{io}(l_i, r_a, r_b)$ equals 1 if there is a resource n that satisfy both (l_i, r_a, n) and (li, rb, n), 0 if not. $C_{ii}(l_i, r_a, r_b)$ equals 1 if there is a resource n that satisfy both (l_i, n, r_a) and (l_i, n, r_b), 0 if not. By extension C_{io} and C_{ii} can be used to compute (1) the total number of indirect and distinct links between r_a and r_b ($C_{io}(n, r_a, r_b)$ and $C_{ii}(n, r_a, r_b)$, respectively outcoming and incoming) as well as (2) the total number of resources n inked indirectly to r_a via l_i ($C_{io}(l_i, r_a, n)$."* (Passant, 2010).

As stated in Petrova et al. (2019a), determination of $LDSD_i$ for recommendations starts as soon as an end user interacts with the system and engages with a building from a result set that was previously returned to a query. Only by implementing the recommender system as such, a real "push" system architecture can be achieved, in which user activity is immediately tracked and recommendations are instantly computed and updated. Using the above described method, the Recommendation Filter component looks for bot:Building objects that are semantically close to each other by relying on all incoming and outgoing links for specific buildings, which are linked in the Building Data Store and the Linkset Store. In other words, the $LDSD_i$ is calculated as a matrix between one building and all semantically close buildings (Petrova et al., 2019a) if the recommender system aims to recommend other buildings. If the recommender system is tailored to recommend alternative buildings based on similarities in operational behaviour, then it might make more sense to compute the semantic distance between buildings using the links between association rules only.

The table of results in Figure 5-9 presents the computed $LDSD_i$ for one of the buildings hosted in the RDF Building Data Store. Since the bot:Building tag is present for all concepts available in the store, it is disregarded. The purpose of that exclusion is to be able to determine semantic relatedness based on the diversity of the bmeta tags. The example used to showcase the approach is, of course limited (six buildings and three different metatags), which also leads to semantic distance values being quite

far apart (0, 0.3333, 0.5 or 1), because only three links are considered: buildingType, designedBy, and energyLabel. The actual knowledge base exposes many more relations and bmeta tags, especially when taking into account the metadata provided through the crowdsourcing tool. In a full implementation, the semantic distances and hence the recommendations will be much more interesting and diverse.

Building	Cio	Cii	LDSD
https://www.example.com/data/buildings/building_2e0dcc1c-b981-4c47-adb4-2b9887f10481	2	0	0.3333
https://www.example.com/data/buildings/building_987d706d-877a-4b1d-80f6-6ee89d856319	1	0	0.5
https://www.example.com/data/buildings/building_43576e80-cf8c-11e1-8000-68a3c4d40f59	1	0	0.5
https://www.example.com/data/buildings/building_af41d889-f50c-456e-9625-96655150838d	0	0	1.0
https://www.example.com/data/buildings/building_aac3427f-eeb0-460c-ba47-14fd44c8be74	0	0	1.0

Figure 5-9: Indirect semantic distances computed for building https://www.example.com/data/buildings/building_00dd6c87-6a6e-f482-7490-e6613659708a (Petrova et al., 2019a)

In terms of recommendations, each retrieved building is also complemented by diverse kinds of data. Such data can be easily retrieved from the full building data graph, which can be enriched as described in Chapter 3 (project data repository). As previously discussed and in line with this thesis, this includes sensor measurements, motifs and association rules discovered in the sensor data, occupant data, etc. Different kinds of metadata and user data can also be displayed, should those be in support of the end user.

Of course, all recommendations and additional data following those need to be displayed in an appropriate end-user interface, which integrates with and supports the BIM-based design processes of the team. Considering the overall framework and system architecture for this thesis discussed in Chapter 2, this user interface and the dedicated recommendations ideally are a part of a CDE.

5.1.3. CHALLENGES AND LIMITATIONS

As with any other system, potential challenges related to the generation of recommendations and the recommender system need to be considered and addressed. One of the most important considerations is related to the user and the user behaviour. The richness of the knowledge base also plays a significant role in the functioning of the recommender system, but the user preferences and their behaviour play just determinative role. As stated in Petrova et al. (2019a), changes of the user profile and preferences over time are essential to the functioning of the system, and therefore have to be continuously evaluated and taken into account, to provide context-aware tailored recommendations. In addition to the changing behaviour, end users may exhibit similar profiles, but different behaviour and preferences depending on the context. Such dynamic behaviour can clearly affect the performance of the recommender system, as the wrong user preferences may be considered by the system. A very important consideration is potential anomalous behaviour such as purposeful negative feedback by the user (Petrova et al., 2019a).

Another limitation may arise from the recommendation approach itself. As stated in (Petrova et al., 2019a), the employed LDSD approach only computes the semantic distance between two resources that are directly or indirectly linked through an intermediate resource and all other resources, which are more than two links away are not considered semantically related. Thus, enhanced LDSD algorithms (Propagated LDSD (Alfarhood et al., 2017)) may need to be used to expand the range beyond the two links distance. Also, the current effort only considers semantic distances between buildings. Semantic distances in accordance with other kinds of metadata may be valuable in the configuration and refinement of the recommender system (Petrova et al., 2019a).

For further details, please refer to Appendix F. Paper VI: "Semantic data mining and linked data for a recommender system in the AEC industry".

CHAPTER 6. CONCLUSIONS

6.1. NEED FOR SUSTAINABILITY IN A WORLD OF CONTINUOUS DIGITAL SHIFTS

Disruptive technologies have an ever increasing impact on society. That also applies to the AEC industry, which finds itself in a continuous redefinition under the influence of digitalisation. The built environment also has a significant contribution to global energy use, climate change, resource depletion, and the well-being of humans. As a result, numerous research efforts exploit the technological advancements in the quest to minimise these negative contributions. These aspects are also reflected in contemporary building design practice, in the core of which lie performance targets aiming to reduce environmental impact and enhance the energy efficiency, indoor environmental quality and comfort for the building occupants.

In that relation, the advent of BIM has caused a paradigm shift in the industry, both in terms of workflow execution and technology adoption. Additionally, the progress in methodological approaches and powerful computational paradigms from the areas of statistical and symbolic AI (e.g. machine learning and semantic data modelling) have made the prediction of design outcomes and the explanation of building performance behaviour possible and much more accurate. Combined with the exponential growth and richness of data generated during the building life cycle, these technologies have the potential to revolutionize the building design process and make it evidence-based.

However, despite these significant technological capabilities and potential, hurdles remain. Fragmentation of the building life cycle, inaccuracy of predictions and design assumptions, poor decision-making and collaboration mechanisms, lack of data integration and sharing across disciplines, and lack of feedback loop from operation to design contribute to the long-standing gap between design intent and measured performance, and discredit the high-performance and sustainable building paradigms.

Thus, this thesis originates in the backdrop of the context of digital transitions and sustainability in the built environment and strives to utilise technology and richness of data as means to enhance human design decision-making with an evidence-based character. The existing building stock contains valuable hidden knowledge disguised both in high-performing and underperforming buildings. The wealth of operational building data and project data repositories can unlock valuable conclusions, which can and should inform future decision-making processes.

More specifically, this research strives to demonstrate how advanced computational methods from the areas of statistical and symbolic AI can be reconciled to help formalize complex engineering knowledge and thereby improve decision-making in

sustainable BIM-based design. Therefore, the main goal of this thesis has been to close the loop between building operation and design to provide evidence-based decision support in a BIM-based sustainable design process.

6.2. SEMANTICS VS. STATISTICS FOR A FEEDBACK LOOP BETWEEN OPERATION AND DESIGN

The main research question that this thesis aims to answer, is formulated as follows:

How can knowledge discovery, representation and retrieval be fused to establish a feedback loop from building operation to design and inform sustainable BIM-based design decision-making in an evidence-based and user-centred way?

This question was originally subdivided in three key objectives:

(1) Provide a framework for performance-oriented design decision support relying on BIM, data mining and semantic data modelling, thereby allowing customized information retrieval according to defined design goals.
(2) Demonstrate how a semantic cloud of building data enriched with performance patterns can be used by design teams as a knowledge base in decision support.
(3) Showcase how the knowledge can be brought back to design professionals through the design aids they use empowered by user-centred context-aware recommendations relying on an ecosystem of rich knowledge bases.

These objectives have been addressed throughout the chapters of this thesis, supported by the collection of papers in appendices A-F. Chapter 1 outlines the main background and challenges in the research area, which leads to the main research question and thesis objectives. Chapter 2 summarises the results of an extensive state of the art review covering the three key research domains, namely KDD, semantic data modelling, and knowledge-based design decision support. Additional literature studies related to supporting subtopics have been appropriately placed throughout the chapters to provide the right context to the presented results. Based on the state of the art review, the main framework for performance-oriented design decision support is developed, which fulfils the first main objective of this thesis.

Chapter 3 effectively implements that framework and thus provides *the first key contribution of this thesis: the system architecture for a framework that can provide user-centred design decision support based on BIM, data mining, and semantic data modelling.* First and foremost, this chapter outlines the different kinds of building data. Furthermore, the chapter demonstrates how each of these data types can be a valuable input for various knowledge discovery or feature matching algorithms, as long as the KDD goal is clearly defined and the data is prepared in accordance with the analytical needs. Semantic data represented with semantic data models and graph models enables reasoning, knowledge representation, disambiguation and querying

(symbolic AI), and numeric data represented with tabular or binary formats enables easy parsing and processing with various machine learning algorithms (statistical AI). The framework proposed in Chapter 3 includes diverse data types, thus enabling the adoption of both symbolic and statistical AI techniques. The framework is tested in two use cases, Gigantium and Home2020, which show how knowledge can be discovered in operational building data, and how the results can be formally represented in a semantic graph to enable information retrieval. As such, knowledge reuse is enabled according to defined design goals, thus implementing user-centred retrieval of discovered knowledge. The results show that knowledge discovery, representation and reuse can be effectively achieved; however, that process is still associated with multiple manual operations and requires a fundamental additional effort: interpretation and contextualisation of the discovered knowledge.

Based on the results from Chapter 3, Chapter 4 proceeds with answering how the resulting semantic graph of building data enriched with performance patterns can be contextualised and made meaningful for the end users (the design team), thus responding to the second objective of this thesis. The results so far show that semantic representation and retrieval of discovered building performance patterns using web technologies is possible; however, many of this data has been stripped of its human-readable meaning. The data is either formalised into its shorter and machine-readable semantic counterpart, or is implicitly present in the form of association rules and motifs that have statistical relevance, but no meaning. Therefore, this part of the thesis presents the *second key contribution, i.e., a crowdsourcing mechanism that allows endowing the available semantic building data enriched with building performance patterns with meaning obtained from human domain experts*. Experts can easily identify those patterns and/or rules that are most likely to be valuable and contain robust hidden knowledge. In accordance with the level of their expertise, domain experts are able to understand the meaning of the patterns in a given context. If these interpretations can be captured, they can create the backbone of a context-aware and semantics-aware decision support system. To achieve this, the thesis employs crowdsourcing techniques and proposes a crowdsourcing network that allows collecting domain expertise for interpretation of association rules in the form of (1) input interpretation; (2) reviews; and (3) upvotes. The validity and feasibility of this system have been demonstrated for one of the use cases previously considered in Chapter 3: Home2020. The crowdsourcing tool allows to effectively add semantic annotations (classifications), human-readable descriptions, as well as votes for association rules in the graph, thereby adding to the full semantic graph and framework that was devised as part of Chapter 3.

With the framework and crowdsourcing tool in place, users can perform all the queries that they may find useful. However, this practically implies the implementation of a pull architecture, in which users are responsible for pulling out the information that they think they need. This does not leave much room for serendipity or surprise, which is incomplete according to the stated objective in terms of context-awareness of the

system. Therefore, Chapter 5 presents *the third key contribution of this thesis, namely a recommender system that brings back knowledge to design professionals through the design aids they use in a user-centred and context-aware manner*. The proposed recommender system complements the overall system architecture (Chapter 3) and the crowdsourcing tool (Chapter 4). The proposed knowledge-based recommender system relies predominantly on the concept of Linked Data Semantic Distance between two concepts, thus resulting in the recommendation of the semantically closest concepts. This approach does not imply that data and recommendations need to be materialised; the recommendations can be computed dynamically, based on input and current status of the full knowledge graph. The recommender system is intended to bring valuable knowledge from the knowledge ecosystem back to the design team through meaningful recommendations based on various levels of similarity with the current context of the design team.

As seen from the summary of contributions, an essential thread in the entire thesis is the fusion of semantic and data analysis (symbolic and statistical) techniques. Building data is represented using semantic data modelling techniques, whereas operational data is analysed using classic data mining techniques; crowd-sourcing is applied in a rigid semantic manner, but also leaves room for the use of analytical techniques with the simple upvoting mechanism; the recommender system relies on semantic techniques, but also adopts analytical techniques in the calculation of semantic distances between concepts. Neither semantics or analytics eventually take the upper hand, but aim to deploy the best of both AI "worlds" to empower the human end user and both technology sets go hand in hand in support of evidence-based sustainable design decision support.

This thesis demonstrates how it is possible to learn from the metabolisms of buildings and their occupants and achieve evidence-based design decision support by fusing symbolic and statistical AI. It was showcased that the finest grains of monitored data can be effectively used to discover performance insights, and use those to build knowledge bases in support of design decision-making. As a result, these techniques can support a revolution in the way buildings are designed, namely by effectively bridging the gap between the operational and design phases in the built environment.

6.3. CHALLENGES AND FUTURE WORK

Besides the presented contributions and the potential identified throughout the research, several challenges were also encountered. Some of them have already been discussed in connection with the results. Each of the presented contributions was delimited to a particular scope, which can be extended to achieve a higher level validity of the results. Some of the challenges, limitations and corresponding recommendations for future research are listed below.

- *Data handling and automation of KDD approaches*- much of the available operational building data, both in this research effort and in general, is

historical data available in logs, even if a real-time data stream from the building is available. The data is usually saved in data lakes and retrieved in batches following the knowledge discovery needs. Regardless of the level of sophistication of the employed knowledge discovery and representation techniques, the discovered knowledge is still a result of batch processing, and does not provide an integrated overview of the performance behaviour of the building beyond the analysed dataset. Even if a direct API link is embedded in the semantic graph, a lot of the associated pre-processing activities are manual and based on extracts from the real-time stream. Future work can thereby consider stream processing technologies, which still rely on the graph structure (e.g. RDF stream processing), but enable pattern discovery directly in the data streams. Preliminary research and experiments presented in Petrova et al. (2019a) show that it is possible to continuously transform the sensor data streams into RDF streams and use semantic data mining techniques on the resulting graph. However, it has to be kept in mind that RDF frequent pattern mining is data structure oriented and based on the graph predicates instead of data values, as opposed to traditional data mining, which focuses only on data values. Future work should, therefore, explore the alignment and consideration of numerical values in the RDF stream.

- *Knowledge interpretation with crowdsourcing techniques*- a key objective discussed at length in this thesis is the interpretation and semantic annotation of the discovered knowledge using crowdsourcing techniques. The initial implementations demonstrated in Chapter 4 testify to the feasibility of the method; however, to fully validate the results, the crowdsourcing platform has to be implemented and tested with domain experts in multiple contexts. The results from the tests will identify potential shortcomings of the approach and provide ideas about the necessary functionality and overall usefulness of the crowdsourcing platform for semantic annotation of building performance patterns. Also, additional studies can be performed here that investigate the potential feasibility of a self-annotating system learning by expert annotation behaviour and how semantic web technologies may be used for assessing the quality and validity of the interpretations and annotations.

- *Linked data-based context-aware recommender system*- similarly to the crowdsourcing effort, the recommender system relying on semantic relatedness between concepts has only been partially implemented in this research effort. To assess the system's usefulness, it has to be fully implemented and tested with design professionals with various profiles and across multiple project contexts. That includes a GUI integrated into the BIM environment for interaction between the design practitioners and the recommender system. Testing the system will provide feedback on several different levels, including usefulness, user engagement, the feasibility of the

recommendation method, size and diversity of the knowledge base, etc. Protocol studies (linkography) can hereby contribute further in the assessment of team dynamics, concept formation and the effect of context on the use of the system. Such a research effort can also contribute to understanding how the design professionals utilise the recommendations.

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, 7(1), 39-59.

Abanda, F.H., Tah, J. & Keivani, R. (2013). Trends in built environment semantic Web applications: Where are we today? *Expert Systems with Applications*, 40, 5563–5577. https://doi.org/10.1016/j.eswa.2013.04.027.

Acosta, M. (2014). Crowdsourcing linked data management. In: A. Bernstein, J. M. Leimeister, N. Noy, C. Sarasua, and E. Simperl (Eds.) *Crowdsourcing and the Semantic Web, Dagstuhl Reports*, 4(7), 29. https://doi.org/10.4230/DagRep.4.7.25.

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S. & Lehmann, J. (2013). Crowdsourcing linked data quality assessment. In: Alani H. et al. (Eds.) *The Semantic Web – ISWC 2013. Lecture Notes in Computer Science*, 8219, 260-276, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41338-4_17.

Abaza, H. (2008). An Interactive Design Advisor for Energy Efficient Buildings. *Journal of Green Building*, 3(1), 112-125. https://doi.org/10.3992/jgb.3.1.112.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, Washington, USA, 207-216. https://doi.org/10.1145/170035.170072.

Ahmad, T., Chen, H., Guo, Y. & Wang, J. (2018). A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. *Energy and Buildings*, 165, 301–320. https://doi.org/10.1016/j.enbuild.2018.01.017.

Ahmad, M.W., Mourshed, M. & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147, 77-89. https://doi.org/10.1016/j.enbuild.2017.04.038.

Ahmed, V., Aziz, Z., Tezel, A. & Riaz, Z. (2018). Challenges and drivers for data mining in the AEC sector. *Engineering, Construction and Architectural Management*, 25, 1436-1453. https://doi.org/10.1108/ECAM-01-2018-0035.

Ahmed, A., Korres, N. E., Ploennigs, J., Elhadi, H., & Menzel, K. (2011). Mining building performance data for energy-efficient operation. *Advanced Engineering Informatics*, 25, 341–354. https://doi.org/10.1016/j.aei.2010.10.002.

Aksamija, A. (2012). *BIM-Based Building performance analysis: Evaluation and simulation of design decisions*. Washington, DC, USA: Omnipress.

Alexander, C. (1977). *A pattern language*. New York, NY, USA: Oxford University Press.

Alfarhood, S., Labille, K. & Gauch, S. (2017). Propagated linked data semantic distance. In: *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, Poznan, Poland, 278-283. https://doi.org/10.1109/WETICE.2017.16.

Alter, S. (2004). A work system view of DSS in its fourth decade. *Decision Support Systems*, 38, 319–327. https://doi.org/10.1016/j.dss.2003.04.00.

Amasyali, K. & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

Arnott, D., & Pervan, G. (2008). Eight key issues for the decision support system discipline. *Decision Support Systems*, 44, 657–672. https://doi.org/10.1016/j.dss.2007.09.003.

Aroyo, L. (2014). Semantic Interpretation and Crowd Truth. In: A. Bernstein, J. M. Leimeister, N. Noy, C. Sarasua, and E. Simperl (Eds.) *Crowdsourcing and the Semantic Web, Dagstuhl Reports*, 4(7), 31. https://doi.org/10.4230/DagRep.4.7.25.

Ashouri, M., Haghighat, F., Fung, B.C., Lazrak, A. & Yoshino, H. (2018). Development of building energy saving advisory: A data mining approach. *Energy and Buildings*, 172 (2018), 139- 151. https://doi.org/10.1016/j.enbuild.2018.04.052.

Ayzenshtadt, V., Langenhan, C., Roth, J., Bukhari, S. S., Althoff, K.-D., Petzold, F., & Dengel, A. (2016). Comparative evaluation of rule-based and case-based retrieval coordination for search of architectural building designs. In: Goel A., Díaz-Agudo M., Roth-Berghofer T. (Eds.) *Case-Based Reasoning Research and Development. ICCBR 2016. Lecture Notes in Computer Science*, 9969, 16-31. https://doi.org/10.1007/978-3-319-47096-2_2.

Baader, F. & W. Nutt (2003). Basic description logics. In *Description logic handbook: theory, implementation, and applications*, 47–100, Cambridge, England: Cambridge University Press.

Barr, A. & Feigenbaum, E. (1981). *The Handbook of Artificial Intelligence*, 1, Los Altos, CA, USA: William Kaufmann, Inc. https://doi.org/10.1016/C2013-0-07690-6.

Beetz, J., van Leeuwen, J & de Vries, B. (2005). An ontology web language notation of the Industry Foundation Classes. In: R. J. Scherer, P. Katranuschkov, & S.-E. Sconfke (Eds.), *Proceedings of the 22nd CIB W78 Conference on Information Technology in Construction*, 193-198, Dresden: Technische Universität Dresden.

Benndorf, G. A., Wystrcil, D. & Rehault, N. (2018). Energy performance optimization in buildings: A review on semantic interoperability, fault detection, and predictive control. *Applied Physics Reviews*, 5, 041501. https://doi.org/10.1063/1.5053110.

Berners-Lee, T., Connolly, D., Kagal, L. & Scharf, Y. (2008). N3logic: a logical framework for the world wide web. *Theory and Practice of Logic Programming*, 8(3), 249–269. https://doi.org/10.1017/S1471068407003213.

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. Scientific American, 284, 34–43.

Beth, E. W. & Piaget, J. (1966). *Mathematical Epistemology and Psychology*. Dordrecht, The Netherlands: Reidel.

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A. & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, 30, 500–521. https://doi.org/10.1016/j.aei.2016.07.001.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Cambridge, United Kingdom: Springer.

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22. https://doi.org/10.4018/jswis.2009081901.

Blohm, I., Zogaj, S., Bretschneider, U., & Leimeister, J.M. (2018). How to Manage Crowdsourcing Platforms Effectively? *California Management Review*, 60(2), 122-149. https://doi.org/10.1177/0008125617738255.

Bonino, D. & Corno, F. (2008). DogOnt - ontology modeling for intelligent domotic environments. In: *Proceedings of the International Semantic Web Conference (ISWC)*, *5318*, *Lecture Notes in Computer Science (LNCS)*, 790–803. https://doi.org/10.1007/978-3-540-88564-151.

Boratto, L., Carta, S., Fenu, G. & Saia, R. (2017). Semantics-aware content-based recommender systems: Design and architecture guidelines. *Neurocomputing*, 254, 79–85. https://doi.org/10.1016/j.neucom.2016.10.079.

Borrmann, A., König, M., Koch, C. & Beetz, J. (2018). *Building Information Modeling: Technology Foundations and Industry Practice*, 1st ed., Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-92862-3.

Brachman, R. J. & Levesque, H. J. (2004). *Knowledge Representation and Reasoning*, Morgan Kauffmann. https://doi.org/10.1016/B978-1-55860-932-7.X5083-3.

Braine, M. D. S. & O'Brien D. P., Eds. (1998). *Mental Logic*. Mahwah, NJ, USA: Erlbaum.

Brickley, D. & Guha, R. (2004). RDF Vocabulary Description Language 1.0: RDF Schema -W3C Recommendation. Retrieved from http://www.w3.org/TR/rdf-schema/

British Standards Institute. (2013). PAS 1192-2:2013 Specification for information management for the capital/delivery phase of construction projects using building information modelling.

Brunato, M. & Battiti, R. (2003). A Location-Dependent Recommender System for the Web. In: *Proceedings of the MobEA Workshop*, Budapest, Hungary.

Burke, R. (2007). Hybrid Recommender Systems, In: P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): *The Adaptive Web. Lecture Notes in Computer Science*, 4321, 377-408. https://doi.org/10.1007/978-3-540-72079-9_12.

Burke, R. (2000). Knowledge-based Recommender Systems, In: A. Kent (ed.): *Encyclopedia of Library and Information Systems*, 69, 32.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Upper Saddle River, NJ, USA: Prentice Hall.

Calbimonte, J.P., Jeung, H., Corcho, O. & Aberer, K. (2012). Enabling query technologies for the semantic sensor web. *International Journal on Semantic Web and Information Systems*, 8, 43–63. https://doi.org/10.4018/jswis.2012010103.

Calbimonte, J.P., Corcho, O. & Gray, A.J.G. (2010). Enabling ontology-based access to streaming data sources. In: *The Semantic Web- ISWC 2010*, 96–111, Berlin, Germany: Springer.

Capozzoli, A., Piscitelli, M., Brandi, S., Grassi, D. & Chicco, G. (2018). Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. *Energy*, 157, 336–352. https://doi.org/10.1016/j.energy.2018.05.127.

Capozzoli, A., Piscitelli, M. S., Gorrino, A., Ballarini, I. & Corrado, V. (2017). Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. *Sustainable Cities and Society*, 35, 191-208. https://doi.org/10.1016/j.scs.2017.07.016.

Capozzoli, A., Serale, G., Piscitelli, M. S. & Grassi, D. (2017a). Data mining for energy analysis of a large data set of flats, In: *Proceedings of the Institution of Civil Engineers. Engineering sustainability*, 1-16. https://doi.org/10.1680/jensu.15.00051.

Capozzoli, A., Grassi, D., Piscitelli, M.S. & Serale, G. (2015). Discovering knowledge from a residential building stock through data mining analysis for engineering sustainability. *Energy Procedia*, 83, 370-379. https://doi.org/10.1016/j.egypro.2015.12.212.

Carbon Trust. (2012). Closing the Gap – Lesson Learned on Realising the Potential of Low Carbon Building Design. London, United Kingdom: Carbon Trust.

Cebrat, K. & Nowak, L. (2018). Revealing the relationships between the energy parameters of single-family buildings with the use of self-organizing maps, *Energy and Buildings*. 178, 61- 70. https://doi.org/10.1016/j.enbuild.2018.08.028.

Chatzikonstantinou, I. & Sariyildiz, I. S. (2017). Addressing design preferences via auto-associative connectionist models: Application in sustainable architectural Façade design. *Automation in Construction*, 83, 108-120. https://doi.org/10.1016/j.autcon.2017.08.007.

Cheng, J.C.P. & Ma, L.J. (2015). A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects. *Building and Environment*, 93(2), 349-361. https://doi.org/10.1016/j.buildenv.2015.07.019.

Chiu, C.M., Liang, T.P. & Turban, E. (2014). What can crowdsourcing do for decision support? *Decision Support Systems*, 65, 40–49. https://doi.org/10.1016/j.dss.2014.05.010.

Consoli, S. & Reforgiato R.D. (2015). An urban fault reporting and management platform for smart cities. In: *Proceedings of the 24th International Conference on World Wide Web*, 535-540. https://doi.org/10.1145/2740908.2743910.

Costa, G. & Madrazo, L. (2014). An information system architecture to create building components catalogues using semantic technologies, In: A. Mahdavi, B. Martens, R. Scherer (Eds.), *Proceedings of the 10th European Conference on Product and Process Modelling (ECPPM)*, 551–557. https://doi.org/10.1201/b17396-90.

Criado-Perez C., Collins, C. G., Jackson, C.J., Oldfield, P., Pollard, B. & Sanders, K. (2019). Beyond an 'informed opinion': evidence-based practice in the built environment. *Architectural Engineering and Design Management*, 1-18. https://doi.org/10.1080/17452007.2019.1617670.

Curry, E., O'Donnell, J., Corry, E., Hasan, S., Keane, M. & O'Rian, S. (2013). Linking building data in the cloud: Integrating cross-domain building data using linked data. *Advanced Engineering Informatics*, 27, 206-219. https://doi.org/10.1016/j.aei.2012.10.00.

Dave, B., Schmitt, G., Faltings, B. & Smith, I. (1994). Case based design in architecture. *Artificial Intelligence in Design- AID '94*, J. Gero and F. Sudweeks (Eds.), 145-162, Dordrecht, The Netherlands: Kluwer Academic Publishers.

de Gemmis, M., Lops, P., Musto, C., Narducci, F. & Semeraro, G. (2015). Semantics-aware content-based recommender systems. In: *Recommender Systems Handbook*, Springer, 119-159.

de Groot, E.H., Mallory-Hill, S.M., van Zutphen, R.H.M. & de Vries, B. (1999). A decision support system for preliminary design. *Durability of Building Materials and Components*, 8, 970-979.

de Souza, C.B. & Tucker, S. (2015). Thermal simulation software outputs: a framework to produce meaningful information for design decision-making. *Journal of Building Performance Simulation*, 8, 57-78. https://doi.org/10.1080/19401493.2013.872191.

de Souza, C.B. & Tucker, S. (2016). Thermal simulation software outputs: a conceptual data model of information presentation for building design decision-

making. *Journal of Building Performance Simulation*, 9(3), 227-254. https://doi.org/10.1080/19401493.2015.1030450.

de Vries, B., Jessurun, J., Segers, N. & Achten, H. (2005). Word graphs in architectural design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 19, 277–288. https://doi.org/10.10170S0890060405050195.

de Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation, *Automation in Construction*, 41, 40 – 49. https://doi.org/10.1016/j.autcon.2014.02.009.

Della Valle, E., Ceri, S., van Harmelen, F. & Fensel, D. (2009). It's a streaming world! Reasoning upon rapidly changing information. *IEEE Intelligent Systems*, 24(6), 83–89. https://doi.org/10.1109/MIS.2009.125.

Delgoshaei, P., Heidarinejad, M. & Austin, M.A. (2018). Combined ontology-driven and machine learning approach to monitoring of building energy consumption. In: *Building Performance Modeling Conference and SimBuild co-organized by ASHRAE and IBPSA-USA*, Chicago, IL, USA, 667-674.

D'Oca, S., Hong, T. & Langevin, J. (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews*, 81, 731 – 742. https://doi.org/10.1016/j.rser.2017.08.019.

D'Oca, S., & Hong, T. (2015). Occupancy schedules learning process through a data mining framework. *Energy and Buildings*, 88, 395–408. https://doi.org/10.1016/j.enbuild.2014.11.065.

Dorst, K., & Cross, N. (2001). Creativity in the design process: Co-evolution of problem-solution. *Design Studies*, 22(5), 425–437. https://doi.org/10.1016/S0142-694X(01)00009-6.

Diaz, J. J.V., Wilby, M.R., Gonzalez, A.B.R., Munoz, J.G. (2013). EEOnt: An ontological model for a unified representation of energy efficiency in buildings. *Energy and Buildings*, 60, 20-27. https://doi.org/10.1016/j.enbuild.2013.01.012.

Dwyer, T. (2013) Knowledge is power: benchmarking and prediction of building energy consumption. *Building Services Engineering Research and Technology*, 4(1), 5–7. https://doi.org/10.1177/0143624412471130.

Elouti, B.H. (2009). Design knowledge recycling using precedent-based analysis and synthesis models. *Design Studies*, 30, 340-368. https://doi.org/10.1016/j.destud.2009.03.001.

El-Diraby, T. E. (2013). Domain ontology for construction knowledge. *Journal of Construction Engineering and Managemen*t, 139(7), 768-784. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000646.

Esnaola-Gonzalez, I., Bermudez, J., Fernandez, I. & Arnaiz, A. (2018). Semantic prediction assistant approach applied to energy efficiency in tertiary buildings, *Semantic Web*, 9, 735-762. https://doi.org/10.3233/SW-180296.

Esnaola-Gonzalez, I., Bermudez J., Fernandez, I. & Arnaiz, A. (2018a). EROSO: Semantic Technologies towards thermal comfort in workplaces, In: *Proceedings of the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, Nancy, France, 519-533. https://doi.org/10.1007/978-3-030-03667-6_33.

Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z. & Wang, J. (2019). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy*, 235, 1551-1560. https://doi.org/10.1016/j.apenergy.2018.11.081.

Fan, C., Song, M., Xiao, F. & Xue, X. (2019a). Discovering Complex Knowledge in Massive Building Operational Data Using Graph Mining for Building Energy Management. *Energy Procedia*, 2481-2487. https://doi.org/10.1016/j.egypro.2019.01.378.

Fan, C., Xiao, F., Li, Z., & Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, 159, 296–308. https://doi.org/10.1016/j.enbuild.2017.11.008.

Fan, C., Sun, Y., Shan, K., Xiao, F. & Wang, J. (2018a). Discovering gradual patterns in building operations for improving building energy efficiency. *Applied Energy*, 224, 116 – 123. https://doi.org/10.1016/j.apenergy.2018.04.118.

Fan, C., Xiao, F., & Yan, C. (2015). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction*, 50, 81–90. https://doi.org/10.1016/j.autcon.2014.12.006.

Fan, C., Xiao, F., Madsen, H., & Wang, D. (2015a). Temporal knowledge discovery in big BAS data for building energy management. *Energy and Buildings*, 109, 75–89. https://doi.org/10.1016/j.enbuild.2015.09.060.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17, 37–54. https://doi.org/10.1609/aimag.v17i3.1230.

Fong, S., Li, J., Song, W., Tian, Y. & Dey, N. (2018). Predicting unusual energy consumption events from smart home sensor network by data stream mining with misclassified recall. *Journal of Ambient Intelligence and Humanized Computing*, 9(4), 1197-1221. https://doi.org/10.1007/s12652-018-0685-7.

Fournier-Viger, P., Wu, C. W., Tseng, V. S., & Nkambou, R. (2012). Mining sequential rules common to several sequences with the window size constraint. In: *Lecture notes in computer science: Vol. 7310. Advances in artificial intelligence*, 299-304, Springer, Berlin, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-30353-1_27.

Fruchter, R., Demian, P., Yin, Z. & Luth, G. (2004). Turning A/E/C knowledge into working knowledge. In: *Towards a Vision for Information Technology in Civil Engineering*, 1-13. https://doi.org/10.1061/40704(2003)12.

Fu, T.C. (2011). A review on time series data mining, *Engineering Applications of Artificial Intelligence*, 24, 164–18. https://doi.org/10.1016/j.engappai.2010.09.007.

Gajzler, M. (2016). Usefulness of mining methods in knowledge source analysis in the construction industry, *Archives of Civil Engineering*, 62, 127-142. https://doi.org/10.1515/ace-2015-0056.

Gajzler, M. (2010). Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry, *Technological and Economic Development of Economy*, 16, 219-232. https://doi.org/10.1515/ace-2015-0056.

Garrett, A. & New, J. (2015). Scalable tuning of building models to hourly data, Energy, 84, 493-502. https://doi.org/10.1016/j.energy.2015.03.014.

Goldman, G., & Zarzycki, A. (2014). Smart buildings/smart(er) designers: BIM and the creative design process. In *Building Information Modeling (BIM) in current and future practices*, 3–16, Hoboken, NJ, USA: Wiley.

Grant, J. & D. Beckett. (2004). RDF test cases. W3C recommendation. Retrieved from http://www.w3.org/TR/rdftestcases/.

Gruber, T.R. (1993). A translation approach to portable ontology specifications, In: *Knowledge Acquisition*, 5(2), 199-220. https://doi.org/10.1006/knac.1993.1008.

Haettenschwiler, P. (2001). Neues anwenderfreundlicheskonzept der entscheidungsunterstützung, vdf Hochschulverlag AG, Zurich, Switzerland, 189-208.

Han, K. & Golparvar-Fard, M. (2017). Crowdsourcing BIM-guided collection of construction material library from site photologs. *Visualization in Engineering*, 5(14). https://doi.org/10.1186/s40327-017-0052-3.

Hall, S., Oldfield, P., Mullins, B. J., Pollard, B. & Criado-Perez, C. (2017). Evidence based practice for the built environment: Can systematic reviews close the research - practice gap? *Procedia Engineering*, 180, 912-924. https://doi.org/10.1016/j.proeng.2017.04.341.

Han, J. W., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques*, 3rd ed., Waltham, USA: Morgan Kaufmann, https://doi.org/10.1016/C2009-0-61819-5

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, USA: MIT Press.

Hayes-Roth, F. & Jacobstein, N. (1994). The state of knowledge-based systems, *Communications of the ACM*, 37(3), 26-39. https://doi.org/10.1145/175247.175249.

Hawkins, J. & Blakeslee, S. (2004). On intelligence. New York, NY, USA: St. Martin's Griffin.

Heylighen, A., Martin, M., & Cavallin, H. (2007). Building stories revisited: Unlocking the knowledge capital of architectural practice. *Architectural Engineering and Design Management*, 3(1), 65-74. https://doi.org/10.1080/17452007.2007.9684630.

Heylighen, A., Neuckermans, H. (2000). DYNAMO: A Dynamic Architectural Memory On-line. *Educational Technology & Society*, (3)2, 86-95.

Heytmeyer, C. L., Pickett, M., Leonard, E.I., Archer, M.M., Ray, I., Aha, D.W. & Trafton, J. G. (2015). Building high assurance human-centric decision systems. *Automated Software Engineering*, 22 (2), 159-197. https://doi.org/10.1007/s10515-014-0157-z.

Hoehndorf, R. & Queralt-Rosinach, N. (2017). Data Science and symbolic AI: Synergies, challenges and opportunities. *Data Science*, 1, 27–38, https://doi.org/10.3233/DS-170004.

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B. & Dean, M. (2004). SWRL: a semantic web rule language combining OWL and RuleML. W3C member submission. Retrieved from http://www.w3.org/Submission/-SWRL/.

Howe, J. (2006). The rise of crowdsourcing, *Wired Magazine*, 14.

Hu, S., Corry, E., Horrigan, M., Hoare, C. M., Reis, D. & O'Donnell, J. (2018). Building performance evaluation using OpenMath and linked data, *Energy and Buildings*, 174, 484-494. https://doi.org/10.1016/j.enbuild.2018.07.007.

Hu, S., Corry, E., Curry, E., Turner, W. & O'Donnell, J. (2016). Building performance optimisation: A hybrid architecture for the integration of contextual information and time-series data, *Automation in Construction*, 70, 51-61. https://doi.org/10.1016/j.autcon.2016.05.018.

International Energy Agency. (2013). Transition to Sustainable Buildings: Strategies and Opportunities to 2050. Retrieved from https://www.iea.org/publications/freepublications/publication/Building2013_free.pdf

Isikdag, U. (2015). *Enhanced building information models: Using IoT services and integration patterns*, 1st ed., Istanbul, Turkey: Springer. https://doi.org/10.1007/978-3-319-21825-0.

Jalaei, F., Jrade, A., & Nassiri, M. (2015). Integrating decision support system (DSS) and building information modeling (BIM) to optimize the selection of sustainable building components, *Journal of Information Technology in Construction (ITcon)*, 20, 399–420.

Janowicz, K., van Harmelen, F., Hendler, J. & Hitzler, P. (2015). Why the data train needs semantic rails. *AI Magazine*, 36(1), 5-14. https://doi.org/10.1609/aimag.v36i1.2560.

Jin, C., Xu, M., Lin, L. & Zhou, X. (2018). Exploring BIM data by graph-based unsupervised learning, In Proceedings of the *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, 582-589. https://doi.org/10.5220/0006715305820589.

Kamari, A., Lausten, C., Peterson, S. & Kirkegaard, P. H. (2018). A BIM-based decision support system for the evaluation of holistic renovation scenarios. *Journal of Information Technology in Construction (ITcon)*, 23, 354-380.

Kim, W. & Katipamula, S. (2018). A review of fault detection and diagnostics methods for building systems, *Science and Technology for the Built Environment*, 24, 3–21. https://doi.org/10.1080/23744731.2017.1318008.

Kim, H., Stumpf, A. & Kim, W. (2011). Analysis of an energy efficient building design through data mining approach, *Automation in Construction*, 20, 37-43. https://doi.org/10.1016/j.autcon.2010.07.006.

Klepeis, N. E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M. & Switzer, P. et al. (2001). The National Human Activity Pattern Survey (NHAPS), Lawrence Berkeley National Lab, CA, United States.

Krijnen, T. & Tamke, M. (2015). Assessing Implicit Knowledge in BIM Models with Machine Learning, In M. Ramsgaard Thomsen, M. Tamke, C. Gengnagel, B. Faircloth, & F. Schreuder (Eds.), *Modelling behaviour. Design Modelling Symposium Copenhagen 2015*, 397-406. https://doi.org/10.1007/978-3-319-24208-8_33.

Lam, K.P., Zhao, J., Ydstie, B.E., Wirick, J., Qi, M. & Park, J. (2014). An EnergyPlus whole building energy model calibration method for office buildings using occupant behavior data mining and empirical data. In: *Proceedings of the 2014 ASHRAE/IBPSA-USA Building Simulation Conference*, Atlanta, GA, USA, 160-167.

Lausch, A., Schmidt, A. & Tischendorf, L. (2015). Data mining and linked open data new perspectives for data analysis in environmental research. *Ecological Modelling*, 295, 5–17. https://doi.org/10.1016/j.ecolmodel.2014.09.018.

Lawson, B. (2005). How designers think- the design process demystified. London, United Kingdom: Architectural Press.

Lefrancois, M., Kalaoja, J., Ghariani, T. & Zimmermann, A. (2017). D2.2: The SEAS Knowledge Model, Technical Report. *ITEA2 12004 Smart Energy Aware Systems*, Brussels, Belgium.

Li, K., Xie, X., Xue, W., Dai, X., Chen, X. & Yang, X. (2018). A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. *Energy and Buildings*, 174, 323-334. https://doi.org/10.1016/j.enbuild.2018.06.017.

Lin, J., Keogh, E., Wei, L. & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15, 107–144. https://doi.org/10.1007/s10618-007-0064-z.

Liu, K. & Golparvar-Fard, M. (2015). Crowdsourcing construction activity analysis from jobsite video streams. *Journal of Construction Engineering and Management*, 141, 04015035. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001010.

Liu, Y., Huang, Y. & Stouffs, R. (2015). Using a data-driven approach to support the design of energy-efficient buildings. *Journal of Information Technology in Construction (ITcon) , Special issue ECPPM 2014-10th European Conference on Product and Process Modelling*, 20, 80–96.

Llanes, K. R., Casanova, M. A., & Lemus, N. M. (2016). From sensor data streams to linked streaming data: a survey of main approaches. *Journal of Information and Data Management*, 7, 130–140.

Magrassi, F., Del Borghi, A. Gallo, M., Strazza, C. & Robba, M. (2016). Optimal Planning of Sustainable Buildings: Integration of Life Cycle Assessment and Optimization in a Decision Support System (DSS). *Energies*, 9, 490. https://doi.org/10.3390/en9070490.

Maher, M. L. & Poon, J. (1996). Modeling Design Exploration as Co-evolution. *Microcomputers in Civil Engineering*, 11(3), 195-209. https://doi.org/10.1111/j.1467-8667.1996.tb00323.x.

Malhotra, M. & Nair, T. R. G. (2015). Evolution of Knowledge Representation and Retrieval Techniques. *Intelligent Systems and Applications*, 7, 18-28. https://doi.org/10.5815/ijisa.2015.07.0.

Manola, F. & E. Miller. (2004). RDF Primer. W3C Recommendation. Retrieved from http://www.w3.org/TR/rdf-primer/.

Mason, K. & Grijalva, S. (2019). A review of reinforcement learning for autonomous building energy management. Retrieved from https://arxiv.org/pdf/1903.05196.pdf.

McGlinn, K., Wagner, A. & Bonsma, P., McNerney, L. & O'Sullivan, D. (2019). Interlinking geospatial and building geometry with existing and developing standards on the web *Automation in Construction*, 103, 235–250. https://doi.org/10.1016/j.autcon.2018.12.026.

McGlinn, K., Yuce, B., Wicaksono, H., Howell, S. & Rezgui, Y. (2017). Usability evaluation of a web-based tool for supporting holistic building energy management, *Automation in Construction*, 84, 154-165. https://doi.org/10.1016/j.autcon.2017.08.033.

McGuinness, D. & van Harmelen, F. (2004). OWL Web Ontology Language - W3C Recommendation. Retrieved from http://www.w3.org/TR/owl-features/

Menezes, C., Cripps, A., Bouchlaghem, D. & Buswell, R. (2012). Predicted vs. actual energy performance of non-domestic buildings: using post-occupancy evaluation data to reduce the performance gap. *Applied Energy*, 97, 355–364. https://doi.org/10.1016/j.apenergy.2011.11.075.

Middleton, S. E., Shadbolt, N. R. & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1), 54–87. https://doi.org/10.1145/963770.963773.

Mihai, A. & Zmeureanu, R. (2013). Calibration of an energy model of a new research center building, In: *Proceedings of the 13th International Conference of the*

International Building Performance Simulation Association (IBPSA), Chambery, France, 1786-1793.

Miller, C., Nagy, Z. & Schlueter, A. (2018) A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*, 81, 1365 – 1377. https://doi.org/10.1016/j.rser.2017.05.124.

Miller, C., Nagy, Z., & Schlueter, A. (2015). Automated daily pattern filtering of measured building performance data. *Automation in Construction*, 49, 1–17. https://doi.org/10.1016/j.autcon.2014.09.004.

Minsky, M. (1991). Logical versus analogical or symbolic versus connectionist or neat versus scruffy, *AI Magazine*, 12(2), 34–51.

Molina-Solana, M., Ros, M., Ruiz, M., Gomez-Romero, J. & Martin-Bautista, M. (2017). Data science for building energy management: A review, *Renewable and Sustainable Energy Reviews*, 70, 598–609. https://doi.org/10.1016/j.rser.2016.11.132.

Musto, C., Basile, P., Lops, P., de Gemmis, M. & Semeraro, G. (2017). Introducing linked open data in graph- based recommender systems. *Information Processing and Management*, 53, 405–435. https://doi.org/10.1016/j.ipm.2016.12.003.

NBIMS-US. (2015). National BIM standard United States version 3, National Institute of Building Sciences, Washington, DC, USA. Retrieved from http://www.nationalbimstandard.org/.

Nilashi, M., Zakaria, R., Ibrahim, O., Majid, M.Z.A., Zin, R.M., Chugtai, M.W., Abidin, N. I. Z., Sahamir, S.R. & Yakubu, D.A. (2015). A knowledge-based expert system for assessing the performance level of green buildings. *Knowledge-Based Systems*, 86, 194–209. https://doi.org/10.1016/j.knosys.2015.06.009.

Ochoa, C.E. & Capeluto, I.G. (2015). Decision methodology for the development of an expert system applied in an adaptable energy retrofit façade system for residential buildings, *Renewable Energy*, 78, 498-508. https://doi.org/10.1016/j.renene.2015.01.036.

O'Donnell, J., Corry, E., Hasan, S., Keane, M. & Curry, E. (2013). Building performance optimization using cross-domain scenario modeling, linked data, and complex event processing. *Building and Environment*, 62, 102-111. https://doi.org/10.1016/j.buildenv.2013.01.019.

Oliveira, J., Delgado, C. & Assaife, A. (2017). A recommendation approach for consuming linked open data. *Expert Systems With Applications*, 72, 407–420. https://doi.org/10.1016/j.eswa.2016.10.037.

Passant, A. (2010). Measuring semantic distance on linking data and using it for resources recommendations. In: *AAAI spring symposium 2010: Linked data meets artificial intelligence*, 93–98.

Pauwels, P., Zhang, S. & Lee, Y.-C. (2017). Semantic web technologies in AEC industry: a literature review. *Automation in Construction*, 73, 145–165. https://doi.org/10.1016/j.autcon.2016.10.003.

Pauwels, P., Krijnen, T., Terkaj, W., & Beetz, J. (2017a). Enhancing the ifcOWL ontology with an alternative representation for geometric data. *Automation in Construction*, 80, 77–94. https://doi.org/10.1016/j.autcon.2017.03.001.

Pauwels, P. & Terkaj, W. (2016). EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology. *Automation in Construction*, 63, 100-133. https://doi.org/10.1016/j.autcon.2015.12.003.

Pauwels, P., de Meyer, R., van Campenhout. (2012). Information system support in construction industry with semantic web technologies and/or autonomous reasoning agents, In: Gudnasson & Scherer (Eds.), *Proceedings of the European Conference on Product and Process Modelling*, Reykjavik, Iceland, 643-653.

Pearson, J. M., & Shim, J. P. (1995). An empirical investigation into DSS structure and environments. *Decision Support Systems*, 13(2), 141–158. https://doi.org/10.1016/0167-9236(93)E0042-C.

Peirce, C. S. (1958). *Collected Papers: Science and philosophy and Reviews, correspondence, and bibliography*, 7, Cambridge, MA, USA: Belknap Press of Harvard University Press.

Pena, M., Biscarri, F., Guerrero, J. I., Monedero, I., & León, C. (2016). Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. *Expert Systems With Applications*, 56, 242–255. https://doi.org/10.1016/j.eswa.2016.03.002.

Peng, Y., Lin, J. R., Zhang, J. P., & Hu, Z. Z. (2017). A hybrid data mining approach on BIM-based building operation and maintenance. *Building and Environment*, 126, 483–495. https://doi.org/10.1016/j.buildenv.2017.09.030.

Perzylo, A., Somani, N., Rickert, M., & Knoll, A. (2015). An ontology for CAD data and geometric constraints as a link between product models and semantic robot task descriptions. In: *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Hamburg, Germany, 4197–4203. https://doi.org/10.1109/IROS.2015.7353971.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019, Under review). Data mining and semantics for decision support in sustainable BIM-based design. Submitted for publication to *Advanced Engineering Informatics.*

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019a). Semantic data mining and linked data for a recommender system in the AEC industry. Accepted for publication in *Proceedings of the 2019 European Conference on Computing in Construction (EC3)*, 10-12 July, Chania, Crete, Greece.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019b, under review). Crowdsourcing building performance patterns for evidence-based decision support in sustainable building design. Submitted to *Automation in Construction*.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data. *Architectural Engineering and Design Management, Special Issue on Intelligent Building Paradigms and Data- Driven Models of Innovation*, 1-23. https://doi.org/10.1080/17452007.2018.1530092.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018a). In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data. In: I. Mutis & T. Hartmann (Eds.) *Advances in Informatics and Computing in Civil and Construction Engineering*, 19-27, Springer. https://doi.org/10.1007/978-3-030-00220-6_3.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018b). From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches. In: J. Karlshøj, & R. Scherer (Eds.) *Proceedings of the 12th European Conference on Product and Process Modelling (ECPPM)*, Copenhagen, Denmark, 391- 399. https://doi.org/10.1201/9780429506215-49.

Piatetsky-Shapiro, G. (1991). Knowledge discovery in real databases: A report on the IJCAI-89 workshop. *AI Magazine*, 11(5). https://doi.org/68–70, 10.1609/aimag.v11i4.873.

Ploennings, J. & Schumann, A. (2017). From Semantic Models to Cognitive Buildings. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, 5105-5106.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago, IL, USA: The University of Chicago Press.

Polanyi, M. (1966). *The tacit dimension*. New York, NY, USA: Doubleday & company.

Pont, U. J., Ghiassi, N., Fenz, F., Heurix, J., Mahdavi, A. (2015). SEMERGY: Application of semantic web technologies in performance-guided building design optimization. *Journal of Information Technology in Construction (ITCon)*, 20, 107-120.

Power, D. J. (2002). *Decision support systems: Concepts and resources for managers*. New York, NY, USA: Greenwood Publishing Group.

Raftery, P., Keane, M. & O'Donnell, J. (2011). Calibrating whole building energy models: An evidence-based methodology. *Energy and Buildings*, 43(9), 2356-2364. https://doi.org/10.1016/j.enbuild.2011.05.020.

Rasmussen, M.H., Lefrancois, M.H., Bonduel, M., Hviid, C.A. & Karlshøj, J. (2018). OPM: An ontology for describing properties that evolve over time. In: *Proceedings*

of the 6th Linked Data in Architecture and Construction Workshop, London, United Kingdom.

Rasmussen, M., Frausing, C., Hviid, C. & Karlshøj, J. (2018). Demo: Integrating Building Information Modeling and sensor observations using semantic web. In: *Proceedings of the 9th International Semantic Sensor Networks Workshop co-located with 17th International Semantic Web Conference (ISWC 2018)*, 48–55, http://ceur-ws.org/Vol-2213/.

Rasmussen, M.H., Pauwels, P., Hviid, C.A. & Karlshøj, J. (2017). Proposing a central AEC ontology that allows for domain specific extensions. In: F. Bosche, I. Brilakis, R. Sacks (Eds.), *Proceedings of the Joint Conference on Computing in Construction*, Heraklion, Crete, Greece, 237-244. https://doi.org/10.24928/JC3-2017/0153.

Resnick, P. & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM*, 40 (3), 56-58. https://doi.org/10.1145/245108.245121.

Richter, K., Heylighen, A. & Donath, D. (2007). Looking back to the future-an updated case base of case-based design tools for architecture. *Knowledge Modelling eCAADe*, 25, 285–292.

Ristoski, P. & Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, 36, 1–22. https://doi.org/10.1016/j.websem.2016.01.001.

Russell, S. & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*, 3rd ed., Upper Saddle River, NJ, USA: Prentice Hall.

Sabri, Q.U., Bayer, J., Ayzenshtadt, V., Bukhari, S.S., Althoff, K.D. & Dengel, A. (2017). Semantic pattern-based retrieval of architectural floor plans with case-based and graph- based searching techniques and their evaluation and visualization. In: *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, Porto, Portugal, 50-60. https://doi.org/10.1016/10.5220/0006112800500060.

Sack, H. (2014). Crowdsourcing for Evaluation and Semantic Annotation. In: A. Bernstein, J. M. Leimeister, N. Noy, C. Sarasua, and E. Simperl (Eds.) *Crowdsourcing and the Semantic Web, Dagstuhl Reports*, 4(7), 43-44. https://doi.org/10.4230/DagRep.4.7.25.

Sacks, R., Lee, C. M. E. G. & Teicholz, P. (2018). *BIM handbook: a guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators*, 3rd ed., Hoboken, NJ, USA: John Wiley & Sons.

Sarasua, C., Simperl. E., Noy, N., Bernstein, A. & Leimeister, J.M. (2015). Crowdsourcing and the Semantic Web: A research manifesto. *Human Computation*, 2(1), 3-17. https://doi.org/10.15346/hc.v2i1.2.

Schenk, E. & Guittard, C. (2011). Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics*, 7, 93–107. https://doi.org/10.3917/jie.007.0093.

Schneider, G.F., Rasmussen, M.H., Bonsma, P., Oraskari, J. & Pauwels, P. (2018). Linked Building Data for Modular Building Information Modelling of a Smart Home. In: *Proceedings of the 12h European Conference on Product and Process Modelling (ECPPM 2018)*, Copenhagen, Denmark, 407–414. https://doi.org/10.1201/9780429506215-51.

Schneider, G.F. (2017). Towards aligning domain ontologies with the Building Topology Ontology. In: *5th Linked Data in Architecture and Construction Workshop*, University of Burgundy, Dijon, France. https://doi.org/10.13140/RG.2.2.21802.52169.

Shen, L., Yan, H., Fan, H., Wu, Y., Zhang, Y. (2017). An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. *Building and Environment*, 124, 388-401. https://doi.org/10.1016/j.buildenv.2017.08.026.

Sheth, A., Henson, C. & Sahoo, S. (2008). Semantic sensor web. *IEEE Internet Computing*, 12, 78–83. https://doi.org/10.1109/MIC.2008.87.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present and future of decision support technology. *Decision Support Systems*, 33, 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7.

Simon, H. A. (1960). The new science of management decision. Upper Saddle River, NJ, USA: Prentice Hall.

Singhaputtangkul, N. & Low, S.P. (2015). Modeling a Decision Support Tool for Buildable and Sustainable Building Envelope Designs. *Buildings*, 5, 521-535. https://doi.org/10.3390/buildings5020521.

Soibelman, L., & Kim, H. (2002). Data preparation process for construction knowledge generation through Knowledge Discovery in Databases. *Journal of Computing in Civil Engineering*, 16(1), 39–48. https://doi.org/10.1061/(ASCE)0887-3801(2002)16:1(39).

Son, H. & Kim, C. (2015). Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches. *Journal of Cleaner Production, Special Issue: Toward a Regenerative Sustainability Paradigm for the Built Environment: from vision to reality*, 109, 144-151. https://doi.org/10.1016/j.jclepro.2014.08.071.

Sowa, J. F. (2008). *Handbook of Knowledge Representation*. Amsterdam, the Netherlands: Elsevier.

Sowa, J. F. (1992). Semantic networks. In: S.C. Shapiro (Ed.), *Encyclopaedia of Artificial Intelligence*, 2nd ed.,1493-1511, New York, NY, USA: John Wiley & Sons.

Sun, C., Zhang, R., Sharples, S., Han, Y. & Zhang, H. (2019). Thermal comfort, occupant control behaviour and performance gap – a study of office buildings in north-east China using data mining. Building and Environment, 149, 305-321. https://doi.org/10.1016/j.buildenv.2018.12.036.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY, USA: Random House.

Szilagyi, I. & Wira, P. (2018). An intelligent system for smart buildings using machine learning and semantic technologies: A hybrid data-knowledge approach. In: IEEE Industrial Cyber-Physical Systems (ICPS), St. Petersburg, Russia, 20-25. https://doi.org/10.1109/ICPHYS.2018.8387631.

Timmermans, H. (2016). Design & decision support systems in architecture and urban planning. In: *Proceedings of the 13th International conference on design & decision support systems in architecture and urban planning*. Netherlands: Springer.

The European Parliament and the Council of the European Union. (2010). Directive 2010/31/EU on the energy performance of buildings. *Official Journal of the European Union*. Retrieved from http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ%3AL%3A2010%3A153%3A0013%3A0035%3AEN%3APDF

Tronchin, L., Manfren, M., & James, P.A. (2018). Linking design and operation performance analysis through model calibration: Parametric assessment on a passive house building. Energy, 165, 26-40. https://doi.org/10.1016/j.energy.2018.09.037.

Tucker, S. & de Souza, C.B. (2016). Thermal simulation outputs: exploring the concept of patterns in design decision-making. *Journal of Building Performance Simulation*, 9, 30–49. https://doi.org/10.1080/19401493.2014.991755.

Turban, E. & Aronson, J.E. (2000). *Decision support systems and intelligent systems*. 6th ed., Upper Saddle River, NJ, USA: Prentice Hall.

Ukkonen, E. (1995). On-line construction of suffix trees. *Algorithmica*, 14(3), 249-260. https://doi.org/10.1007/BF01206331.

van Leeuwen, J. P & Timmermans, H. J. P. (2004). *Recent advances in design and decision support systems in architecture and urban planning*. Dodrecht, The Netherlands: Kluwer Academic Publishers.

Wang, A. & Srinivasan, R. (2016). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75, 796-808. https://doi.org/10.1016/j.rser.2016.10.079.

Wang, X., Zhang, X. & Li, M. (2015). A survey on semantic sensor web: Sensor ontology, mapping and query. *International Journal of u- and e- Service, Science and Technology*, 8, 325–342. https://doi.org/10.14257/ijunesst.2015.8.10.32.

Weber M., Langenhan C., Roth-Berghofer T., Liwicki M., Dengel A. & Petzold F. (2010). a.SCatch: Semantic structure for architectural floor plan retrieval. *Case-Based Reasoning. Research and Development. Lecture Notes in Computer Science*, 6176, 510-524, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14274-1_37.

Weiner, P. (1973). Linear pattern matching algorithms. In: *The 14th Annual IEEE Symposium on Switching and Automata Theory*, USA, 1–11. https://doi.org/10.1109/SWAT.1973.13.

Wolf, S., Møller, J. K., Bitsch, M. A., Krogstie, J. & Madsen, H. (2019). A Markov-switching model for building occupant activity estimation. *Energy and Buildings*, 183, 672-683, https://doi.org/10.1016/j.enbuild.2018.11.041

Xiang, W., Sun, L., You, W. & Yang, C. (2018). Crowdsourcing intelligent design. *Frontiers of Information Technology & Electronic Engineering*, 19(1), 126-138. https://doi.org/10.1631/FITEE.1700810.

Xiao, X., Skitmore, M. & Hu, X. (2017). Case-based reasoning and text mining for green building decision-making. *Energy Procedia*, 111, 417-425. https://doi.org/10.1016/j.egypro.2017.03.203.

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, 75, 109–118. https://doi.org/10.1016/j.enbuild.2014.02.005.

Xin, H., Meng, R. & Chen, L. (2018). Subjective Knowledge Base Construction Powered By Crowdsourcing and Knowledge Base. In: *SIGMOD'18: 2018 International Conference on Management of Data*, Houston, TX, USA, 1349-1361. https://doi.org/10.1145/3183713.3183732.

Yang, Y. & Hongyang, L.T. (2018). A social networking enabled crowdsourcing system for integrated infrastructure asset management. In: Proceedings of the *Construction Research Congress 2018*, 322-330. https://doi.org/10.1061/9780784481295.033.

Yarmohammadi, S., Pourabolghasem, R. & Castro-Lacouture, D. (2017). Mining implicit 3D modeling patterns from unstructured temporal BIM log text data. *Automation in Construction*, 81, 17- 24. https://doi.org/10.1016/j.autcon.2017.04.012.

Yu, Z. J., Haghighat, F. & Fung, B. C. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. *Sustainable Cities and Society* , 25, 33-38. https://doi.org/10.1016/j.scs.2015.12.001.

Yu, Z., Fung, B., & Haghighat, F. (2013). Extracting knowledge from building-related data- A data mining framework. *Building Simulation*, 6(2), 207–222. https://doi.org/10.1007/s12273-013-0117-8.

Zero Carbon Hub. (2010). A Review of the Modelling Tools and Assumptions: Topic 4, Closing the Gap between Designed and Built Performance, London, United Kingdom: Zero Carbon Hub.

Zhang, Y., Bai, X., Mills, F.P. & Pezzey, J.C.V. (2018). Rethinking the role of occupant behavior in building energy performance: A review. *Energy & Buildings*, 172, 279–294. https://doi.org/10.1016/j.enbuild.2018.05.017.

Zhang, C., Cao, L. & Romagnoli, A. (2018a). On the feature engineering of building energy data mining. *Sustainable Cities and Society*, 39, 508-518. https://doi.org/10.1016/j.scs.2018.02.016.

Zhang, C., Beetz, J., & de Vries, B. (2018b). BimSPARQL: Domain-specific functional SPARQL extensions for querying RDF building data. *Semantic Web*, 9(6), 829–855. https://doi.org/10.3233/SW-180297.

Zhao, H. X. & Magoules, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16, 3586-3592. https://doi.org/10.1016/j.rser.2012.02.049.

Zhong, B., Gan, C., Luo, H. & Xing, X. (2018). Ontology-based framework for building environmental monitoring and compliance checking under BIM environment. *Building and Environment*, 14, 127-142. https://doi.org/10.1016/j.buildenv.2018.05.046.

Zhu, N., Anagnostopoulos, A. & Chatzigiannakis, I. (2018). On mining IoT data for evaluating the operation of public educational buildings. In: *2018 IEEE International Conference on Pervasive Computing and Communications Workshops*, Athens, Greece, 278–283. https://doi.org/10.1109/PERCOMW.2018.8480226.

PUBLICATIONS FOR THE THESIS

Thesis Title: AI for BIM-based sustainable building design: Integrating knowledge discovery and semantic data modelling for evidence-based design decision support

Name of the Ph.D. Student: Ekaterina Aleksandrova Petrova

Name of Supervisor: Associate Professor Kjeld Svidt

Name of Assistant Supervisor: Associate Professor Rasmus Lund Jensen

List of Publications:

A) Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data. *Architectural Engineering and Design Management, Special Issue on Intelligent Building Paradigms and Data- Driven Models of Innovation*, 1-23. https://doi.org/10.1080/17452007.2018.1530092

B) Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data. In: I. Mutis & T. Hartmann (Eds.) *Advances in Informatics and Computing in Civil and Construction Engineering*, 19-27, Springer. https://doi.org/10.1007/978-3-030-00220-6_3

C) Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019, under review). Data mining and semantics for decision support in sustainable BIM-based design. Submitted to *Advanced Engineering Informatics.*

D) Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches. In: J. Karlshøj, & R. Scherer (Eds.) *eWork and eBusiness in Architecture, Engineering and Construction: Proceedings of the 12th European Conference on Product and Process Modelling (ECPPM)*, Copenhagen, Denmark, 391- 399. London: CRC Press/Balkema. https://doi.org/10.1201/9780429506215-49

E) Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019b, under review). Crowdsourcing building performance patterns for evidence-based decision support in sustainable building design. Submitted to *Automation in Construction.*

F) Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019a, in press). Semantic data mining and linked data for a recommender system in the AEC industry. Accepted for publication in *Proceedings of the 2019 European*

Conference on Computing in Construction, 10-12 July, Chania, Crete, Greece.

Other publications (not included in the thesis):

G) Petrova, E. & Pauwels, P. (2019c, under review). Pattern ReCognition in Sustainable Architectural Design: Assessing the Effects of Context and Team Dynamics with Protocol Studies. Submitted to *Research in Engineering Design.*

H) Petrova, E., Romanska, I., Stamenov, M, Svidt, K. & Jensen, R.L. (2017). Development of an Information Delivery Manual for Early Stage BIM-based Energy Performance Assessment and Code Compliance as a Part of DGNB Pre-Certification. In: *Proceedings of the 15th IBPSA conference, Building Simulation 2017*, San Francisco, USA, 2100-2109. https://doi.org/10.26868/25222708.2017.556

I) Petrova, E., Johansen, P.L., Jensen, R.L., Maagaard, S. & Svidt, K. (2017). Automation of Geometry Input for Building Code Compliance Check. In: F. Bosché, I. Brilakis & R. Sacks (Eds.) *LC3 2017 : Volume I – Proceedings of the Joint Conference on Computing in Construction*, Heraklion, Crete, Greece, pp.617-626. https://doi.org/10.24928/JC3-2017/0265

J) Petrova, E., Rasmussen, M., Jensen, R.L. & Svidt, K. (2017). Integrating Virtual Reality and BIM for end-user Involvement in Building Design: a case study. In: F. Bosché, I. Brilakis & R. Sacks (Eds.) *LC3 2017 : Volume I – Proceedings of the Joint Conference on Computing in Construction*, Heraklion, Crete, Greece, pp. 699-709. https://doi.org/10.24928/JC3-2017/0266

K) Petrova, E. (2018). Let the data tell you the truth. Data-driven decision support for high-performance building design. *H V A C Magasinet*, *54*(6), pp. 28-33.

This thesis has been submitted for assessment in partial fulfilment of the Ph.D. degree. The thesis is based on the submitted and published scientific papers in appendices A-F listed above. Parts of the papers have been used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to both the assessment committee and the Faculty of Engineering and Science.

APPENDICES

Appendix A. Paper I

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data. *Architectural Engineering and Design Management*, *Special Issue on Intelligent Building Paradigms and Data- Driven Models of Innovation*, pp. 1-23.

https://doi.org/10.1080/17452007.2018.1530092

Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data

Ekaterina Petrova[a], Pieter Pauwels[b], Kjeld Svidt[a], and Rasmus Lund Jensen[a]

[a]*Department of Civil Engineering, Aalborg University, Aalborg, Denmark*
[b]*Department of Architecture and Urban Planning, Ghent University, Ghent, Belgium*

Sustainable building design requires an interplay between multidisciplinary input and fulfilment of diverse criteria to align into one high-performing whole. BIM has already brought a profound change in that direction, by allowing execution of efficient collaborative workflows. However, design decision-making still relies heavily on rules of thumb and previous experiences, and not on sound evidence. To improve the design process and effectively build towards a sustainable future, we need to rely on the multiplicity of data available from our existing building stock. The objective of this research is, therefore, to transform existing data, discover new knowledge and inform future design decision-making in an evidence-based manner. This article looks specifically into this task by (1) outlining and distinguishing between the diverse building data sources and types, (2) indicating how the data can be analysed, (3) demonstrating how the discovered knowledge can be implemented in a semantic integration layer and (4) how it can be brought back to design professionals through the design aids they use. We, therefore, propose a performance-oriented design decision support system, relying on BIM, data mining and semantic data modelling, thereby allowing customised information retrieval according to a defined goal.

Keywords: BIM, Sustainability, Building Design, Semantics, Data Mining, Pattern Recognition, Knowledge Discovery, Information Retrieval

Introduction

Sustainable building design requires an optimal interplay between diverse criteria, susceptible to both the fulfilment of strictly formulated requirements, as well as their interpretation, translation and implementation by the design team. Hence, a performance-oriented design process requires multidisciplinary input to align into one high-performing 'whole', simultaneously with that being done in the most efficient way. 'Whole' as a concept, and the derived term 'holism', was defined by Smuts (1926) as *'a unity of parts, which is so close and intense as to be more than the sum of its parts'*. That means that all parts should function towards the whole, determine each other and eventually merge their individual characters, which makes the holistic character

This is an Accepted Manuscript of an article published by Taylor & Francis in Architectural Engineering and Design Management, Special Issue on Intelligent Building Paradigms and Data-Driven Models of Innovation, available online: http://dx.doi.org/10.1080/17452007.2018.1530092.

discoverable in the functions of both the parts and the whole. This concept is translated into whole building design by the implementation of the integrated design approach. Therefore, sustainable design requires a holistic approach, in which there are no individual parts constituting a design, only synergetic multidisciplinary inputs that contribute to the targeted overall performance of the whole.

In that relation, Building Information Modelling (BIM) (Eastman et al., 2011; Sacks et al., 2018) has already brought a profound change to the Architecture, Engineering and Construction (AEC) industry by allowing much more efficient integrated workflows. Open data standards and protocols, including Information Delivery Manuals (IDMs), Model View Definitions (MVDs), Industry Foundation Classes (IFC), etc. (buildingSMART, 2016) have served as catalysts towards increased collaboration between stakeholders. This is crucial for obtaining efficiency gains and successful fulfilling of performance targets related to sustainability in the building design domain. By definition, BIM allows integration of multidisciplinary information within a single coordinated building model and empowers collaborative practices (Zanni et al., 2017).

Furthermore, BIM practice strongly advises the use of a Common Data Environment (CDE) to manage information from all stakeholders. The CDE is defined as '*a central repository where construction project information is housed. The contents of the CDE are not limited to assets created in a 'BIM environment' and it will therefore include documentation, graphical model and non-graphical assets.'* (British Standards Institute, 2013). In a CDE, distinct viewpoints on a building are brought together, thus providing the place where a holistic view is possible. That includes data that is often not captured directly in a BIM model (e.g. design briefs, point cloud data, etc.) (Fig. 1).

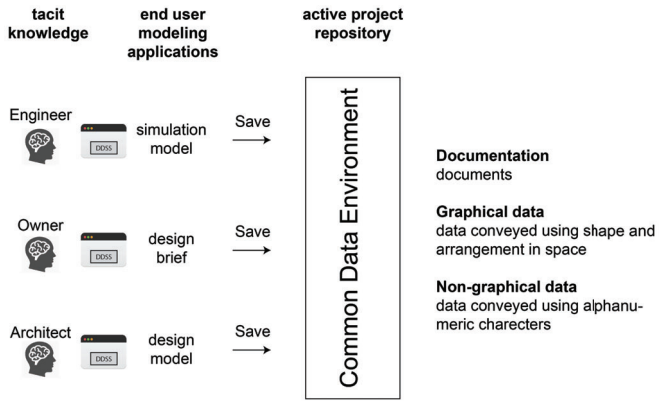

Figure 1. Use of a Common Data Environment in collaborative building design

As a result of the strong focus on BIM, BIM-based sustainable design has received major attention, and is a part of fundamental research within the construction

industry (Cemesova et al., 2015; Lu et al., 2017; Wong & Zhou, 2015). A considerable research effort, aiming for the seamless integration of BIM and building performance assessment in the (early) design process has also taken place in the last decade (El-Diraby et al., 2017; Ilhan & Yaman, 2016; Jalaei & Jrade, 2014; Liu et al., 2015; Schlueter & Thesseling, 2009; Shadram et al., 2016; Underwood & Isikdag, 2010; Yalcinkaya & Singh, 2015).

Even though BIM offers possibilities for synergy with sustainable design, many of the decisions taken during the design process are based on rules of thumb and previous experiences (Heylighen et al., 2007), which are not directly applicable or are not based on sound evidence. Polanyi (1958) defines such rules of thumb and experiences as tacit knowledge, and indicates that it is hard to capture, formalize and make explicit because of its context-specific nature. The increase in experience leads to more complex rules of thumb, which evolve into design patterns (Alexander, 1977). These patterns are crucial in one's understanding of what constitutes and satisfies the design context and heavily influence the design process.

Nevertheless, knowledge discovered in data from past projects and buildings in operation can be combined with the tacit knowledge for informing future design decision-making. As a result, huge potential would arise in achieving building design in a sustainable, efficient and evidence-based manner. One of the main research objectives in this regard is to leverage the multiplicity of data sources and types, and thus pave the way to knowledge discovery for evidence-based processes in design and engineering practice. To advance towards achievement of this objective, this study aims to employ the latest advances in three main areas:

(1) the full use of BIM as a means to reuse existing project data (e.g. through a CDE),

(2) the deployment of Knowledge Discovery in Databases (KDD) (Fayyad, 1996) to discover hidden knowledge in operational building data and inform future building design decision-making, and

(3) the reliance on semantic data modelling to represent the discovered knowledge in a semantically rich graph of data.

Despite not being the main focus, we hereby aim to also take into account the tacit knowledge and expertise used in design decision-making. The main principle is to identify meaningful and relevant patterns from previous projects and buildings in operation, transform information, discover new knowledge and better predict outcomes. The discovered knowledge will provide the basis for a design decision support system (DDSS), which is performance- and data-informed, rather than just data-dependent. Decision support systems are regarded here as computer-based tools adapted to support and aid complex decision-making and problem solving (Arnott & Pervan, 2008; Shim et al., 2002). Research in this area typically highlights the importance of information technology in improving the efficiency and effectiveness of decision-makers (Alter, 2004; Pearson & Shim, 1995). In the context of architectural design and engineering,

research limits more specifically to DDSS targeting the end user (Timmermans, 2016). Many commercial tools (CAD tools, BIM tools, simulation, visualization and coordination tools, etc.) have also been widely adopted in practice. However, they are most often stand-alone applications that do not implement the concept of knowledge reuse. We therefore aim to bring those features together in a DDSS that enables both knowledge sharing and reuse.

Methodological approach

This research relies on an extensive literature review aiming to identify both seminal works and state of the art developments within multiple research areas. Included here are design thinking and theory, BIM, sustainable building design and performance assessment, data analysis and artificial intelligence in performance-oriented architecture and civil engineering, as well as emerging technologies and computational approaches for improvement of design decision-making. We hereby also try to take into account design workflows in various settings. Based on this background research, we investigate the existing types of building data, their representations, formats, storage methods, and the way in which they can be handled by various algorithms, relative to variable goals of the knowledge discovery processes.

Next, we devise a system architecture that aims to bring the knowledge discovered in the available data to the end user and thereby support decision-making in future performance-oriented design processes. This system relies on three main approaches targeting knowledge discovery, namely data mining, geometric feature matching and direct semantic queries. We investigate to what extent the results of geometric similarity matching and data mining can be represented in semantic graphs, thereby relying on earlier work (Petrova et al., 2018a, 2018b). The resulting framework would therefore be able to successfully combine these approaches in support of AEC domain specialists working towards improving the built environment.

In this article, we first document key efforts for information exchange and data analysis in sustainable building design (Section 2). Section 3 proposes a system outline for holistic sustainable design relying on operational building data and project data repositories. Sections 4 and 5 summarize the proposed system, thereby indicating the main implementation methods, i.e. data mining, geometric feature matching and direct semantic queries. Finally, Section 6 presents a conclusion and outlines future work.

Data Exchange and Analysis in Collaborative Sustainable Building Design

Data-Driven and Experience-Based Design

Sustainability is a multi-dimensional matter, aiming for equal balance between economic and social development, and environmental protection (United Nations, 2010). From a collaborative perspective, Senciuc et al. (2015) define sustainable design as a complex system of elements linked by interdependencies and a process of

managing numerous perspectives. Furthermore, Kocaturk (2017) underlines the important role that technology plays in transforming the understanding of sustainability as a concept in the built environment, by enabling design innovation at product, process and operational levels. Sonetti et al. (2018) further highlight the potential of artificial intelligence and ICT tools for human-centric regenerative design. Building performance, on the other hand, besides being a criterion itself, is an outcome of a multidisciplinary set of multiple-criteria design decisions (Jalaei, et al., 2015). In that relation, the availability of data and the efficiency of its exchange are highly influential to both the design decision-making and its results. However, building design is characterized by fragmentation of processes and heterogeneity of actors, competencies and information sources. As a result, data is not readily available and not necessarily easily exchanged. As stated by Akin (2014), the information created and associated with the design must be available and applicable at all stages, without any losses, duplication of trivial processes or backtracking.

According to Aksamija (2012), high-performance design requires *"building performance predictions, use of simulations and modelling, research-based and data-driven processes."* BIM can facilitate knowledge transfer and experience between ongoing projects, but it is also important to use the experience from previous projects to adopt a holistic standpoint (Goldman & Zarzycki, 2014). Thus, for the design intent and performance targets to be achieved, the building operation needs to inform the design, and both phases should not be considered separate or independent, but parts of a cause and effect relationship. Furthermore, Goldman & Zarzycki (2014) claim that much of the data initially required for modelling could be based on predictions relying on data from previous projects. That would require pairing substantial data collection with captured professional expertise. Yet, the result would be a refined outcome, where quantified knowledge and professional experience are used in decision-making in a dedicated and structured way. According to Isikdag (2015), such a future transformation needs a *"focus on enabling an (i) integrated environment of (ii) distributed information which is always (iii) up to date and open for (iv) derivation of new information."* Goldman & Zarzycki (2014) further stipulate that a future data exchange network also has to be based on reuse of experience across designers, and requires knowledge to be modular and shareable.

Basics of Data Analytics and Application of KDD in the AEC Industry

Data analysis is becoming increasingly important for the built environment. Through the emergence of BIM, information as a concept has paved the way to changing the way professionals in the industry work. However, many questions still need to be answered with regards to what should be measured, how the information should be reported and stored, and most importantly, how it should be translated to knowledge and applied in practice. In that relation, Starkey & Garvin (2013) take a step back and highlight the variable, sometimes intertwining definitions of the terms data, information and knowledge from philosophical, semiotic and cybernetic points of view. From a

knowledge management perspective, Thierauf (1999) defines data as "unstructured facts and figures that have the least impact". Davenport & Prusak (2000) claim that, for data to become information, it needs to be contextualised, categorised, calculated and condensed, whereas knowledge implies know-how, meaning and understanding.

This article adopts the term data in a foundational way, as the building blocks for information, which in turn allows purposeful pattern discovery in various datasets, by the use of dedicated analytical approaches. The obtained analytical results would further allow combinations in support of cognitive processes in design. More specifically, the term 'data' in the current context refers to various types and representations of digital data, generated and available throughout the entire building life cycle. That includes generated design documentation (design brief databases) graphical design data (BIM models, simulation models, numeric geometric data), and non-graphical data (semantic design data, numeric simulation output, monitored operational performance data from sensor networks), etc. In other words, we refer to digital building data types in representations useful for further computational analyses. We explicitly focus on digital data and its representations to reflect and comply with the BIM and CDE-based workflows. The article further highlights the potential impact that discovered applicable knowledge in digital data can have on the future built environment.

From an analytical perspective, large volumes of data prove to be overwhelming when using traditional methods, which generate informative reports, but fail when it comes to analysis of their content (Soibelman & Kim, 2002). On the other hand, data mining, KDD and pattern recognition excel at the analysis of data and extraction of knowledge, and can facilitate an effective design space exploration.

Hand et al. (2001) define data mining as *"the analysis of large observational datasets to find unsuspected relationships and summarize the data in novel ways so that data owners can fully understand and make use of the data."* Additionally, Bishop (2006) states that *'pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories'*. In that context, Piatetsky-Shapiro (1991) formulates knowledge as the end product of a data-driven discovery, whereas KDD represents the overall process of the extraction of useful knowledge. Data mining is the step in that process which employs specific algorithms to discover useful and previously unknown patterns in the data. Fayyad et al. (1996) state that the essential purpose is to discover high-level knowledge in low-level data. Furthermore, they define five essential steps, which transform the available raw data into actionable knowledge and insights of immediate value to the end user (Fig. 2).

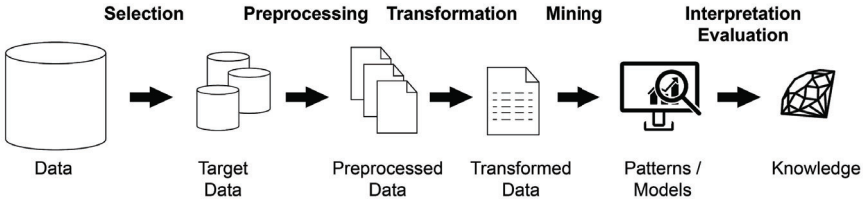

Figure 2. Knowledge Discovery in Databases (KDD) process, Fayyad et al. (1996)

(1) Selection

Data selection deals with the necessity to develop and understand the application domain, capture the relevant prior knowledge and identify the goal of the KDD process from an end-user perspective. Thereafter, a suitable target dataset or subset of variables should be chosen.

(2) Pre-processing

Pre-processing includes cleansing of the data in terms of handling of missing data fields, removal of duplicates, as well as fusion and resolution of conflicts due to the data originating from heterogeneous sources. Soibelman & Kim (2002) argue the significant importance of data preparation to the generation of high-quality knowledge through KDD. In addition, Cabena et al. (1998) point out that 60% of the time goes into data preparation, whereas the mining itself accounts for only 10% of the overall effort.

(3) Transformation

Transformation is concerned with reduction and projection of data with the purpose of finding useful features and representing the data according to the needs of the stated goal and the chosen algorithms. That includes finding invariant data representations and using dimensionality reduction methods to reduce the effective number of considered variables.

(4) Data mining

Data mining deals with matching the defined KDD goals with a particular method, e.g. classification, regression, or clustering. That includes the selection of algorithms and pattern extraction methods, as well as considerations concerning the end user's capabilities for interpretation of the chosen model vs. the model's predictive capabilities and accuracy. The actual data mining can then take place, i.e. searching for patterns in a particular representational form or set of representations, such as rule sets, trees, clusters, etc.

(5) Interpretation / Evaluation

The last step involves interpretation of the mined patterns and examination of their validity. That may include visualization of the discovered patterns and

assessment of their usefulness. Of particular importance is acting on the discovered knowledge, e.g. documenting it, using it directly, or implementing it into another system for further use.

Related Works

Fayyad et al. (1996) define six widely accepted data mining categories, namely classification, clustering, association rule mining, regression, summarization and anomaly detection. Han et al. (2012) further detail each of these techniques and highlight their belonging to two main categories: predictive (supervised) and descriptive (unsupervised). Supervised techniques are powerful for predictive modelling and knowledge representations (regression or classification models). They describe the qualitative or quantitative relationships between the input and output variables, and rely on domain expertise and training data (a set of observations, for which both the input and output variables are given). Thus, discovery of novel knowledge with predictive techniques is therefore unlikely, because inputs and outputs are predefined.

Unsupervised techniques (e.g. clustering, association rule mining, etc.), on the other hand, hold a significant potential in discovering the intrinsic structure, correlations and associations in data. Training data has no relation to the success of unsupervised analytics, as inputs and outputs are not predefined. In that relation, Han et al. (2012) state that the fundamental advantage of unsupervised methods lies within the ability to discover previously unknown and hidden knowledge in the given data. Unlike supervised approaches that adopt a backward approach by having a predefined target, unsupervised analytics are forward oriented, which gives the possibility of discovering interesting relationships and bringing out the value in the data (Fan et al., 2018).

As a result of their potential, KDD and data mining approaches have received major attention in the AEC industry. We performed a literature review that identifies main areas of application in the context of sustainability and energy efficiency, both from predictive and descriptive perspectives. Predictive applications include building energy use and demand prediction (Ahmed et al., 2011; Wang & Srinivasan, 2017; Zhao & Magoulès, 2012), prediction of building occupancy and occupant behaviour (D'Oca & Hong, 2014; Zhao et al., 2014), and fault detection diagnostics for building systems (Cheng et al., 2016; Pena et al., 2016). Descriptive tasks, on the other hand, are concerned with framework development (D'Oca & Hong, 2015; Fan et al., 2015a, 2015b; Park et al., 2016; Yu et al., 2013; Zhou et al., 2015), patterns in occupant behaviour (Capozzoli et al., 2017), building modelling and optimal control (Xiao & Fan, 2014), as well as discovering and understanding energy use patterns (Gaitani et al., 2010; Miller et al., 2015; Wu and Clements-Croome, 2007). Other efforts include the use of data mining for high-performance building design based on classification models for sustainability certification evaluation (Jun & Cheng, 2017), use of BIM-based data mining approaches for improvement of facility management (Peng et al., 2017), use of semantic modelling, neural networks and data mining algorithms for building energy management (McGlinn et al., 2017), etc.

However, the use of KDD and pattern recognition has been dedicated mostly to improvement of the building operation. Using discovered knowledge to improve future building design processes is an area that is rarely explored in detail. Efforts include pattern recognition in simulation data and extraction of information from BIM design log files (Yarmohammadi et al., 2016), use of data-driven approaches to design energy-efficient buildings by mining of BIM data (Liu et al., 2015) and data mining for extracting and recommending architectural design concepts (Mirakhorli et al., 2015).

Reuse of similarities for design decision support has also been recognised in design practice. This is prominent in case-based reasoning (CBR), which provides decision makers with a problem solving framework involving recalling and reusing previous knowledge and experience (Aamodt & Plaza, 1994). CBR approaches in design differ based on the method of their implementation (Elouti, 2009; Heylighen & Neuckermans, 2000; Richter et al., 2007). Example implementations in the context of sustainable architectural design can be found in (Sabri et al., 2017; Shen et al., 2017; Xiao et al., 2017).

In addition, research targeting the creation of a "repository of knowledge" for decision support based on patterns in thermal simulation output has been significantly extended in de Souza & Tucker (2015), de Souza & Tucker (2016) and Tucker & de Souza (2016). All similarity retrieval efforts mentioned above occupy the same conceptual space and are of high relevance to this research. Yet, despite coming a step closer to realizing the targeted future process, they rely on patterns only in design and simulation data. Thus, we aim to contribute further by adopting the latest semantic technologies, adding operational data mining and geometry matching capacities, and taking into account BIM and CDE-based workflows in early design.

The data analysis results coming from existing buildings and designs can rarely be linked to an early stage design using computational tools, mainly because the data representations do not match. This is not the case for tacit knowledge, which facilitates intuitive associations to any visual representation in an early design stage. A design professional would therefore tend to rely primarily on that knowledge instead of tangible performance data. In terms of data analysis, traditional approaches typically start from the available data and focus on retrieving the inherent insights. Decision-makers then determine how these insights may help them. As a result, despite the importance of the KDD goal definition, the knowledge discovery is driven only to a limited extent by the needs of the decision-maker.

Advanced analytical approaches start from the decision-makers and the identification of the most critical decisions, including the variability of their potential outcomes. As a result, the necessary insights to clarify those decisions can be identified, the type of information they may stem from, the data sources that could provide this information, and the knowledge to extract. Thus, a more user-oriented analysis is targeted, resulting in useful and practically applicable design decision support.

Towards Holistic Sustainable Design Relying on Operational Building Data and BIM Data Repositories

The ultimate objective of this research effort is to propose a DDSS that can bring forward a much more efficient sustainable building design process. More specifically, we aim to achieve informed decision-making by reusing existing BIM data repositories and operational building data. BIM data can include BIM models, simulation data, design briefs, etc.; operational data includes monitored data from existing buildings, i.e. sensor data, building use data, and so forth. The purpose is to integrate the DDSS in both the CDE as well as the individual end-user applications. That is found necessary, as the CDE hosts the information related to the building design process, and the end user applications host the individual decisions.

Data and Knowledge with Potential Impact on Design Decision Support

When implementing an advanced data analytics approach, there are several considerations, pertaining not only to the goals and criticality of the decisions, but also to the ability to generalize over the discovered patterns. Meaningful patterns are those that can be statistically justified, hence they should be based on the exploration of significant volumes of heterogeneous data. Furthermore, such an approach has highest impact when it can affect both the design process and the final product. In summary, the suggested approach works best in an environment that hosts simultaneously:

- decisions with high impact and criticality, namely early-stage design decisions with high level of variability of outcome
- specific performance criteria, concerning the practical implications of the decisions with regards to targeted building performance
- data from a high number of reference buildings
- data in big amounts and diversity

Many of the critical early decisions and the related requirements and constraints are interdependent. These dependencies can be captured in diagrams, which give a full overview of the relevant decision-making criteria and relations. Predictive models can hereby contribute further, by quantifying the weights of the dependencies, the criticality of the decisions, the variability of outcomes and the potential impacts. Figure 3 shows the developed dependency diagram capturing the relevant decision-making criteria in high-performance design. The grey nodes with most dependencies highlight not only the criticality of the related decisions, but also the data that would be most relevant for goal-oriented analytics. AEC projects generate various kinds of data in different formats, however, not all data are equally useful to all pattern recognition techniques. The following sections categorize the diverse data types based on their origin.

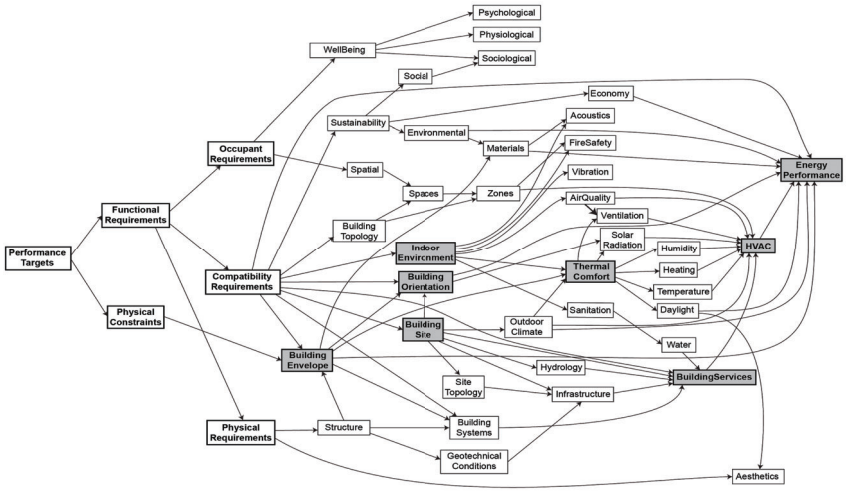

Figure 3. Criteria dependency in a typical sustainable design process

Data Types and Hidden Knowledge at Building Operation Stage

Operational building data is usually represented in a two-dimensional structured tabular way, with columns representing variables and rows storing the measurements at given time steps. Collected data usually includes time and date of measurement, energy consumption data (e.g. power consumption, cooling and heating loads, etc.), HVAC system operating conditions (temperature, flow rates, etc.), and environmental data (e.g. indoor and outdoor climate, humidity, solar radiation, etc.). These data types consist of parameters that are directly influencing building performance and are dynamically changing. Such data are a valuable input for data-driven simulations, HVAC system optimization and improvement of the building operation. Figure 4 represents the dynamic parameters and therefore operational data types typically collected from Building Management Systems (BMS).

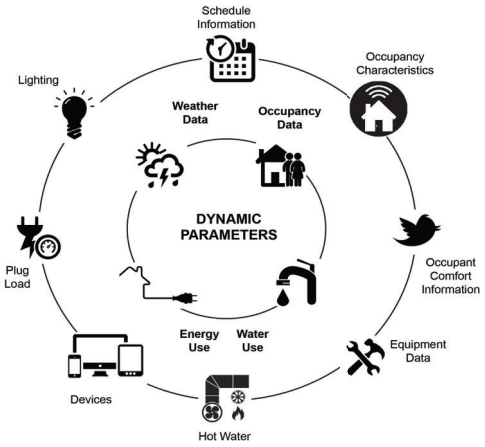

Figure 4. Dynamic parameters, based on taxonomy by Mantha et al. (2015)

According to Han et al. (2015), the typical formats and the tabular representation of operational building data gives an opportunity for discovery of two main types of knowledge: cross-sectional (static) and temporal (dynamic). Cross-sectional knowledge can be discovered when treating each row as an independent observation. The discovered knowledge is static, as the temporal dependencies between the rows are ignored (the knowledge discovered mainly includes the concurrent relationships among the different variables). Static knowledge discovery is useful for the identification of interaction between system components, atypicality in operation, etc. Han et al. (2015) further state that, in contrast, temporal knowledge can be discovered by mining data along both axes of the two-dimensional table and is very useful for characterizing dynamics in building operations. The insights obtained can be used for developing dynamic solutions for optimal building control, fault detection and diagnosis. Capturing the temporal dependencies in the data are much more challenging, but give a possibility for discovering unsuspected patterns and their relationships.

Data Types and Hidden Knowledge at Building Design Stage

The knowledge discovered in design data is much more static, even when taking into account versioning possibilities. Data at the building design stage typically starts with a design brief and a design model. Crucial choices on building orientation, zoning, spatial arrangement, and building materials are made in the earliest design stages. This data typically responds to the requirements and constraints listed earlier in the dependency diagram in Fig. 3 and represents important static parameters defining the character of the building (Fig. 5).

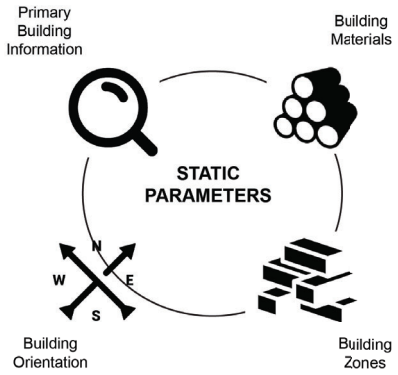

Figure 5. Static parameters, based on taxonomy by Mantha et al. (2015)

A lot of hidden knowledge is also available in the simulation data. This data can inform the design according to the paths defined in the dependency diagram by giving an insight into the building performance. Yet, they are typically a lot more optimistic compared to the actual performance. Building geometry is also valuable, as it provides many of the inputs required for simulation and compliance checking.

Data Type Definition from Analytical Perspective

To achieve high success rate in terms of analytical evaluation, it is important to match the types of data with the most suitable analytical techniques. Different data types can be recognized, informing the choice of analytical techniques and the structure of the data to enable effective knowledge discovery and performance-oriented decision support. The list below presents a data type definition from an analytical perspective.

- Semantic design data: semantic data describing design features, which include building elements, materials, object types, design brief data, etc.
- Numeric geometric data: geometric data in a format optimized for geometric analysis.
- Numeric sensor data: tabular sensor data with real-time data from supervisory control and data acquisition systems.
- Numeric simulation data: data models containing simulation results.

A Holistic Approach to a Data-Driven Sustainable Design System

This section proposes a system architecture that combines the available data with data analytics in a sensible way for decision support. This analysis is put forward through Fig. 6, which shows the main approach and the overall flow of proposed activities.

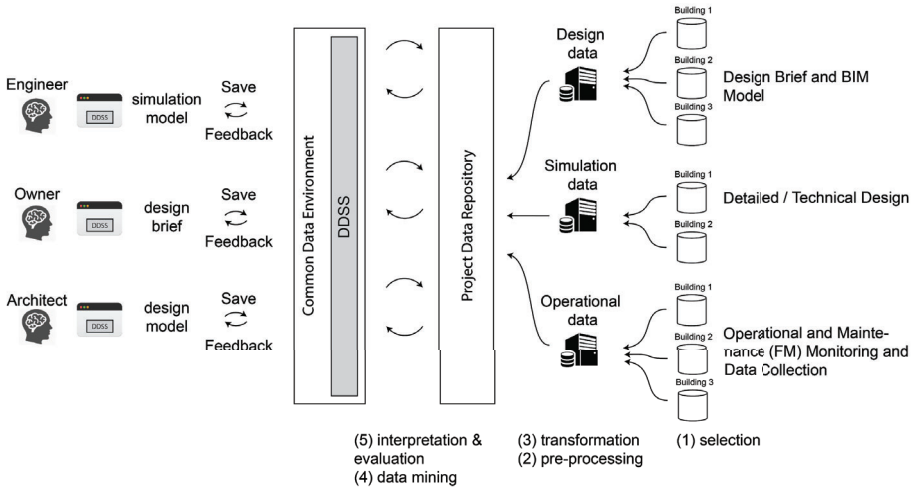

Figure 6. Proposed flow of data from existing buildings and project data repositories towards the diverse end-users

The active design environment (left in Fig. 6) may include BIM authoring tools, parametric design tools, simulation tools, etc. Design professionals iterate through a number of proposals within their individual tools and with the rest of the team. While designing, project data is stored in the CDE as files being uploaded to a central server.

In this study, DDSS systems are proposed both in the CDE and in the individual applications, where the DDSS in the CDE communicates to a project repository (Fig. 6). This repository collects the data available from previous projects and existing buildings, which comes from various heterogeneous sources. For example, BIM data captures the design, but typically comes in different representations, including a native 3D model, a neutral IFC data model, schedules, etc. Sensor data comes in different representations, depending on the system from which it originates. Storing local copies facilitates the execution of the data selection part of the KDD process defined by Fayyad et al. (1996) together with the maintaining of the original data. The selected data can then be cleansed and transformed, thus following steps 2 and 3 by Fayyad et al. (1996). After cleansing and transformation of the selected datasets, the results are stored in a project data repository, which hosts disparate data. While this allows diverse analysis techniques, integration across the data types will be needed.

The following sections indicate how the different components of the proposed system can be set up. We focus specifically on how different approaches may be effectively combined to achieve useful design decision support. Section 4 deals with the part of the system architecture related to the active design environment, including the semantic integration of data, while Section 5 introduces the use of KDD for creating a project data repository.

The Active Design Environment

End-users approach decision-making in an iterative problem-solution oriented manner, in which they put forward solutions based on tacit knowledge. When it comes to the DDSS, an insight into the cognitive processes within design decision-making provides an invaluable input for system design. We therefore first consider the overall design thinking processes, after which we outline how this takes form in a BIM-based process that relies on a CDE with heterogeneous data.

Design Thinking and Problem Solving as a part of Data-Driven Design

The background knowledge of the decision-maker determines the course of the design process. With each design iteration, designers explore a problem/solution space, thereby going through a continuous co-evolution of problem and solution (Dorst & Cross, 2001). As already indicated, the digital part of this process typically happens in a CDE, which stores the multidisciplinary design solutions as they come in sequentially. All actors go through a co-evolution process using their own tacit background knowledge and technology stack. The design requirements, typically captured in the design brief, drive the design decisions and follow the co-evolution of problem and solution. In the context of sustainable design, both the tacit definition for sustainable design and the solution responding to the particular requirements evolve throughout the design process. Ideally, the design team converges over time, under the influence of the design brief and the performance targets, both in the problem and solution spaces (Fig. 7). Convergence brings the team closer to a solution that fulfils the targets. The purpose is to avoid regress, e.g. widening of cycles at any given point in the evolution of the time dimension.

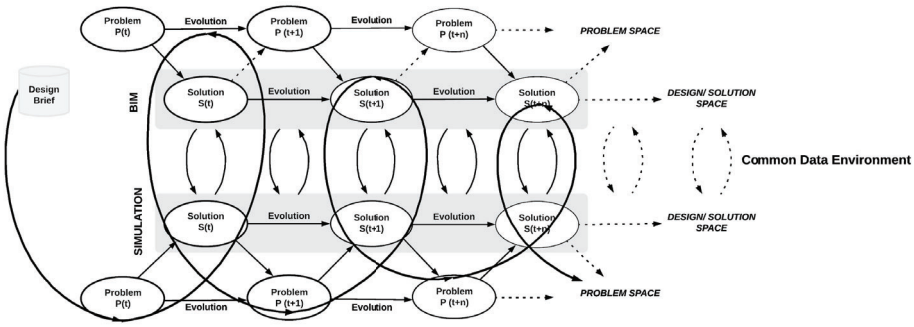

Figure 7. Problem-Solution cycle in collaborative design

In order to give tangible performance data a better role in the above process, the way in which decision-makers connect to their own background knowledge needs to be influenced. This can only be done by presenting the decision-maker useful alternatives

(problem-solution space), which match the goal and build on the tacit experience in a structured way.

Tools and Data Flows in the Active Design Environment

Even if a CDE is used, data is typically kept in separate files. This makes an integrated view over the available information very difficult to achieve. More recent initiatives aim at making the data available in an integrated manner using web technologies. As the web is evolving into a web of data instead of a web of documents (Berners-Lee et al., 2001), technology can be used to make the CDE web-compliant and data-oriented, as opposed to its current document-based nature. Such a system is much more attractive as (1) it makes project data available for semantic information retrieval and management, (2) it allows a larger diversity of data mining approaches, as data can be processed multiple times for different purposes while maintaining the same semantic identifiers, and (3) advanced semantic data mining techniques are within scope. Building a web-based semantic CDE results in the design environment outlined in Fig. 8.

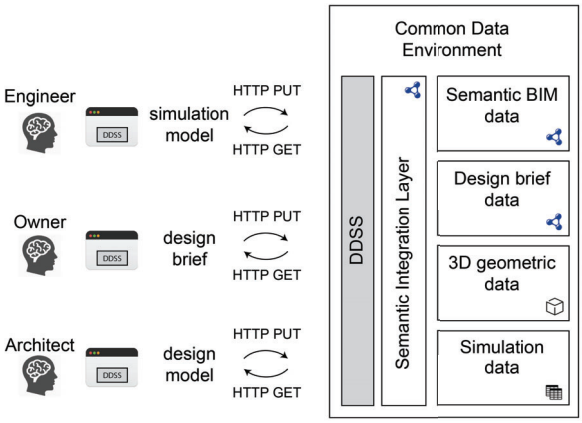

Figure 8. Integration of datasets in a web-based CDE

As the CDE has a web-based structure, applications and users are less occupied with manually storing files in an online server. Instead, the CDE is automatically filled with data using the HTTP protocol. By doing so, a lot more versioning and data logging can be achieved. Considering that data is gathered from multiple heterogeneous sources, the CDE would function optimally with a decentralized structure, which is most commonly realized using graph database approaches. Promising solutions in this regard for the AEC domain relate to deployment of linked data and semantic web technologies (Pauwels et al., 2017a). These technologies allow to build a decentralized web of semantic information, which serves perfectly for maintaining the backbone of a web-based CDE, thereby allowing to link the diverse datasets together, while respecting their original data structures.

Research has also shown that not all data can be efficiently maintained in a graph database or triple store (Pauwels et al., 2017b). We suggest that vast amounts of numeric data, such as geometric, simulation, and sensor data are therefore explicitly kept out of the semantic graph. Geometric data, such as 3D meshes, 2D drawings, point cloud data, etc., are ideally maintained in formats that can efficiently be parsed by geometric analysis algorithms. Sensor and simulation data are typically stored in tabular formats. Therefore, we propose a semantic integration layer (Fig. 8), which maintains the links between the individual datasets. The semantic integration layer is a thin and modular structure, capturing the key semantics of the different data sources in a decentralized manner, while referring to the original data sources that are kept in their optimized structures. The CDE can then be used to query the project data repository.

Reusing BIM Project Data and Operational Building Data

Matching queries from the CDE with the project data repository can occur in a number of ways, depending on how the data is stored. In this section, we look into the structure of the project data repository, and how pattern recognition and matching techniques can be applied to the data (direct queries, geometric feature matching, data mining). An overview diagram of the project data repository is given in Fig. 9.

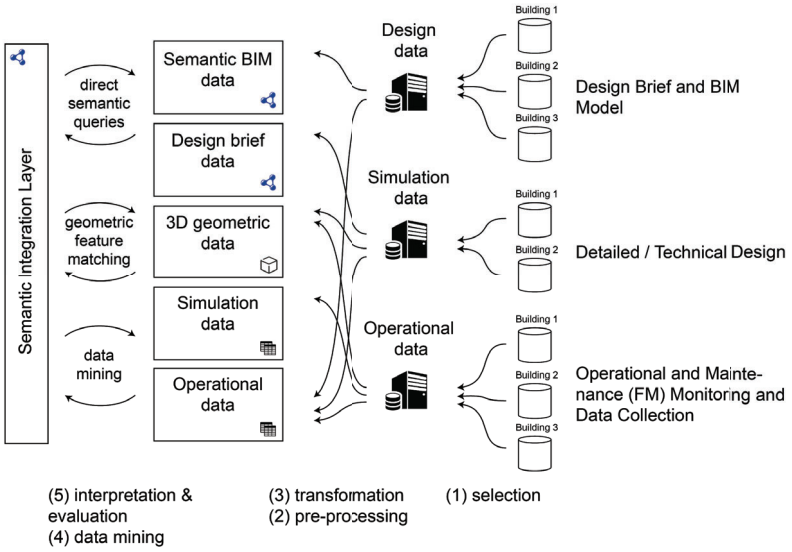

Figure 9. Overview of the project data repository

Structure of the Project Data Repository

Although a project data repository does not necessarily need to have the exact same structure as the CDE, it should be similarly well-structured. By maintaining this data

structure, and not converting all data into linked data, for example, we aim to allow as many as possible feature matching and data mining algorithms. Indeed, it is possible to transform all data to a semantic format, and then to query this data directly (Ristoski & Paulheim, 2016). Yet, this would disallow many of the efficient data mining and geometry matching algorithms that can be used for retrieving knowledge. Instead, we propose to store the semantic, geometric, and operational data separately. These datasets are then interlinked through the semantic data integration layer, which aims to link the semantic data model of a building with its numeric forms.

Clearly, the sole reliance on direct semantic information retrieval queries will be insufficient to give full feedback to an end user targeting a holistic performance-oriented design. The semantic queries do not capture the diversity of conclusions and matches that can be gathered from data mining techniques. Furthermore, relying solely on data mining techniques will not provide the integrated view over the diverse datasets. The same applies to geometric data; one cannot rely only on geometric data to retrieve valuable knowledge from a project repository to inform a designer aiming at holistic sustainable design solutions. Therefore, the diverse data sources need to be available and dynamically linked to allow information retrieval and design decision support.

To build a project data repository as proposed, a number of crucial steps need to be made. Data needs to be selected, cleansed and transformed so that it fits the project data repository. Furthermore, it is advisable to prepare separate local copies of the data in order not to intrude or violate data integrity at the source. In the selection process, it is possible to select only the data of relevance and place them on a local server (see step 1 in Fig. 9). For the static data, such as a design model, design brief, and simulation data, a direct copy can be used. For the dynamic data, such as the operational data and sensor data, data streams need to be accessed continuously. By implementing this data selection process, not only is the data in scope, but the original data is also maintained secure. In a next stage, data can be cleansed and transformed (steps 2 and 3 in Fig. 9). These are highly necessary steps to allow data mining with accurate results. The main purpose of the data transformation step is to end up in the structured project data repository as outlined above.

Recognizing Patterns from the Hive

Data Mining for Temporal Knowledge Discovery in Operational Building Data

Operational building data updates continuously with additional data points. The result is a data stream that gives an indication of the building operation (the heartbeat of the building). The dynamics in operation are usually very complex, due to changes in outdoor climate, indoor occupancy, systems utilization, etc., which rarely occur simultaneously. Discovering related temporal knowledge is of valuable importance to decision-making concerning building components, building automation and control systems, etc. Fan et al. (2015a) state that operational data is in essence multivariate time series data, where each observation is a vector of multiple measurements and control

signals, and time intervals between subsequent observations are usually fixed. That means that using temporal knowledge discovery can help capture relationships between variables over a particular time period.

Various approaches have been developed for temporal knowledge discovery of patterns, e.g. events, clusters, motifs (frequent sequential patterns), discords (infrequent sequential patterns) and temporal association rules, but rarely in the context of operational building data. A framework developed by Fan et al. (2015a) demonstrates encouraging potential in temporal knowledge discovery for improvement of building operations and performance management.

To inform design decision-making, it is important that the discovered patterns hold the potential to increase the confidence of the decisions, while still allowing creativity and variability of design space exploration. Considering the target data in this case and the goal for discovery of unsuspected patterns and relationships, unsupervised temporal knowledge mining should target motifs (and/or discords), as well as association rules (Fu, 2011). Motifs are by themselves valuable to temporal association rule mining and discord detection. We propose to use multivariate motif discovery as a first step (Vahdatpour et al., 2009), as it gives the possibility to discover both synchronous and asynchronous multivariate motifs consisting of univariate motifs or subsets of motifs. That is important, as in this context, motifs in operational building data do not necessarily start at the same time or have the same length. For example, turning the air conditioner on does not lead to an immediate change in indoor temperature due to the thermal mass (Fan et al., 2015a). Employing this method makes it possible to first discover univariate motifs and then use graph clustering approaches to identify multivariate motifs.

In addition, association rule mining (ARM) can help discover associations between variables (Agrawal et al., 1993). ARM usually targets cross-sectional knowledge and temporal dependencies are neglected. Due to the complexity and dynamics of operational building data, the use of temporal association rule mining (TARM) would be more useful, because it provides not only an insight into the associations between the variables, but also their temporal dependencies (Fournier-Viger et al., 2012). As a result, applying the above-mentioned techniques will allow decision-making support by identifying complex patterns over time, as well as the dependencies in their occurrence.

Feature Matching in Geometric Data

Geometric data can also be used for matching data in the CDE with data in the project data repository. Direct geometric pattern matching techniques can be implemented and used to return the most resembling results to a user. A number of geometry types and representations can be considered. One of the most commonly used is IFC, which is a neutral data model aiming to capture building semantics and object properties along with the full 3D geometry. IFC provides one of the most expressive neutral data models to describe building geometry in full semantic detail. A number of

alternative open data models are available as well. One example is the geometry ontology defined by Perzylo et al. (2015). Furthermore, Well-Known Text (WKT)[1] is a markup language that also allows specifying geometry with simple strings based on common agreement. Most WKT content refers to 2D geometry and is used for geospatial data, but it could also be used for representing 3D building geometry (Pauwels et al., 2017b).

Most of the above geometric data models can be captured in the form of labelled graphs. Yet, geometric topology graphs are slightly different, as they typically focus on the nodes and edges representing lines, boundaries, and points. An example of such a geometric topology graph is given for a room with four walls in Fig. 10.

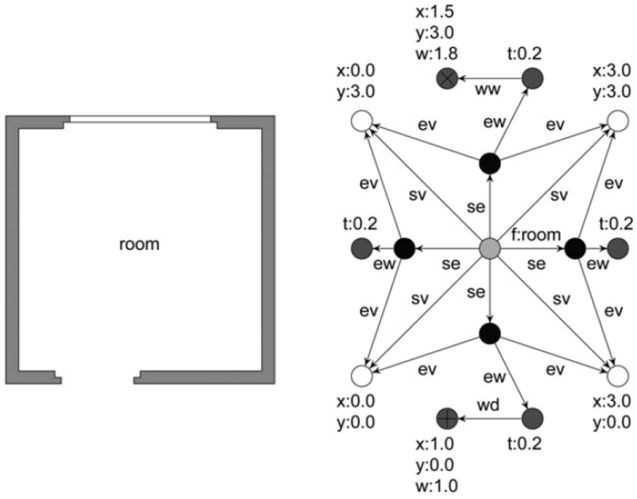

Figure 10. Geometric topology graph, Strobbe et al. (2016)

3D building data can also be represented using 3D mesh models. Yet, such data is semantically less defined and direct geometric feature matching techniques are less applicable. Point cloud data are also used to represent geometry, but, similarly to 3D mesh data, this data structure presents limited semantics.

For semantically rich geometric models, graph matching techniques can be used. Several direct graph matching techniques are available, in particular in data-oriented or web-oriented contexts. SPARQL, CYPHER, and GraphQL are graph query languages used for graph matching in a CDE. This technique assumes the target data to be available in graphs, which can be the case for IFC, WKT, and geometric topology models, but not for the rest.

Advanced geometric analysis algorithms can work with semantically unspecific data, such as point cloud data or 3D meshes, in order to make sense of the unstructured data and match them with the current geometric data in the CDE. Geometric analysis algorithms aim at parsing input geometry, including the unstructured mesh and point

[1] http://www.opengeospatial.org/standards/wkt-crs

cloud data. These are typically hardcoded algorithms, able to evaluate geometry and distil specific characteristics. The extracted characteristics are typically semantic and can thus be captured in a semantic data structure. Examples here are the GeoSPARQL[2] and BimSPARQL (Zhang et al., 2018) query languages, the first aiming at geospatial data and the second aiming at building data. The query languages contain statements such as "within" and "above", thus allowing to formulate geometric semantic queries.

Direct Semantic Queries

Another way to match data from a CDE to a project data repository is through direct semantic queries. Such queries can target the semantic integration layer, the semantic design model data and/or the semantic attributes that may be inferred from data mining or geometric feature recognition techniques.

The modular ontology structure proposed by the W3C Linked Building Data (LBD) Community Group[3] can serve to capture the considered semantics in an efficient way. This includes a number of ontologies, such as a Building Topology Ontology (BOT) (Rasmussen et al., 2017), a PRODUCT ontology, a PROPS ontology (properties), and an Ontology for Property Management (OPM). These ontologies allow to represent the building topology, product data, element properties and management of those properties. The OPM ontology is specifically useful, as it captures desired property values and whether they are achieved or not. Recent industry implementations further target the representation of design brief requirements in commercial graph databases, such as Neo4J, which is highly similar to the linked data approach. Hence, a semantically rich graph is possible based on OPM, BOT, PRODUCT, and PROPS ontologies.

Using linked data technologies, links can be maintained with the operational and geometric data. Device data can be captured using SAREF[4], home automation data can be represented using DogOnt (Bonino & Corno, 2008), and aggregate sensor data can be represented using SSN[5] and/or SOSA[6]. However, these ontologies do not serve well in case all operational data are targeted. In such case, a tabular format is still a lot more effective. The mentioned ontologies can be used to capture static characteristics, such as averages, min-max values, features of interest, devices, etc. The results of the geometric analysis algorithms can be captured in semantic graphs. These are static semantic annotations added to the semantic graph. Full geometric matching is however best done using the original data in a non-semantic format.

The semantic integration layer makes the connection with the non-semantic data possible, namely the reference source for operational data (web server address of

[2] http://www.opengeospatial.org/standards/geosparql
[3] https://www.w3.org/community/lbd/
[4] https://w3id.org/saref
[5] https://www.w3.org/TR/vocab-ssn/
[6] https://www.w3.org/ns/sosa/

specific sensor node data) and geometric data (web server address of specific geometric data file). The integration layer connects the semantic, geometric and operational data, so that any system accessing the data can recognize the associations.

Proposed System Architecture

The proposed system architecture utilizes measured operational building data and project data, which then serve as an input for the discovery of useful knowledge by the use of selected goal-oriented pattern recognition algorithms. The top in Fig. 11 represents the active design environment, which communicates with the project data repository (bottom in Fig. 11). This repository collects all reference data, linked together using the semantic integration layer, but also kept in their native formats. It is enriched using direct semantic queries, geometric feature matching, and data mining techniques, thereby allowing data-driven decision support for holistic performance-oriented design.

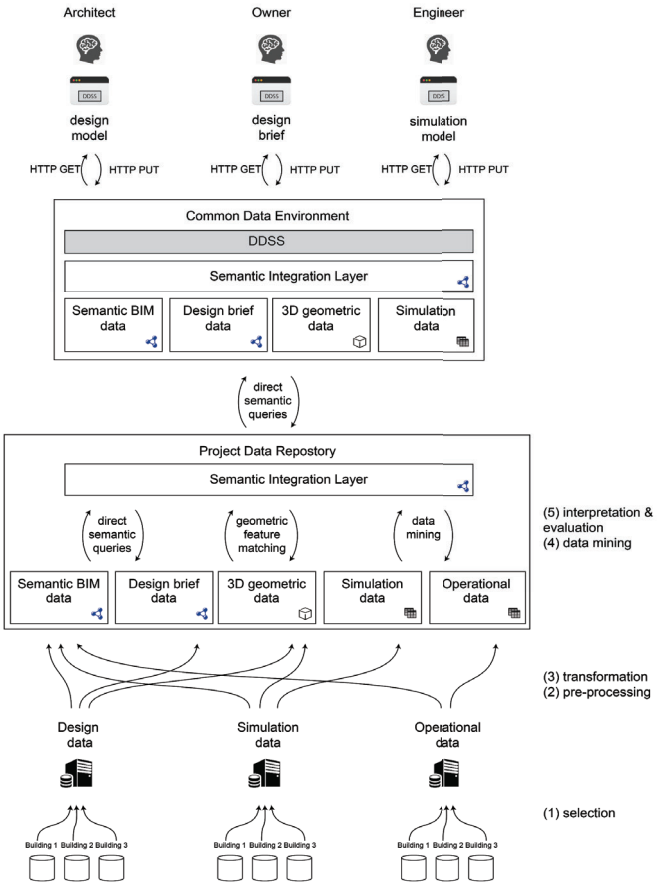

Figure 11. Proposed system architecture

Conclusion

This paper presents a framework for data-driven performance-oriented building design, relying on decision support from knowledge discovered in operational building data and project data repositories. The work identifies the relevant data types and combines three main approaches targeting knowledge discovery accordingly, namely data mining, geometric feature matching and direct semantic queries. The research identifies that the outcome of both the geometric similarity matching and the data mining can be represented in semantic graphs, which allows building a decision support system employing direct semantic queries. The combined approach allows semantic integration of heterogeneous datasets, their attributes and instances. The user-defined semantic queries allow customised information retrieval according to a defined goal.

One of the key challenges identified in this work is the implementation of a semantic integration layer, which combines data from various sources in a semantic graph, yet still allows to deploy data mining and geometric feature matching techniques. Although it is possible to include explicit results from these approaches in a graph (Petrova et al., 2018b), this might compromise the flexibility and modularity of the DDSS. By deploying the proposed web-based system architecture, we hope to overcome this challenge and make the data analysis and information retrieval user-driven. Such approach aims to integrate, yet also preserve the multiplicity of data and algorithms, allowing to deploy them to the maximum of their capabilities, in support of holistic sustainable design.

Future work needs to be done with regards to the testing and implementation of the proposed system in environments that can respond to the necessary requirements: design decisions with high impact and criticality, specific performance criteria, high number of reference buildings, and access to data in big amounts and diversity. Considering the diverse data analysis algorithms and web-based information retrieval approaches, the practical implementation needs to happen in an incremental and modular fashion, ideally involving a community knowledgeable in the architectural design, engineering and construction domains. This implementation process will indicate necessary changes in terms of performance, practical applicability, etc.

More importantly, however, this implementation process needs to reflect and capture the direct value that can be obtained in each concrete stakeholder environment. Of critical importance in future research are the methods that are used to ‘match past and present’ (CDE and project data repository). This match has not been discussed here at length. Choosing which matching mechanism (data mining, direct semantic queries, geometric feature matching) is used when, is of critical importance for the functioning of the system and needs to be investigated in further detail.

The proposed framework can be of significant importance for collaborative design teams aiming to improve the quality of the built environment in terms of sustainability, energy performance, indoor environmental quality, HVAC system design, etc. That includes a number of scenarios and contexts. This research effort targets the early design phase, where the decisions have the biggest impact on the future

performance. Thus, matching needs to be done as early as possible in the design process. The early design phase is, however, also one of the most difficult phases to provide decision support, because of the very limited amount of specific information that is available at this stage. Data is usually limited to an overall site definition, a design brief, and a preliminary layout of spaces. Most designers initially work in a 3D modelling environment, performing mass studies and spatial design exploration. Little semantic information can be obtained in such tools in contrast to the detailed data that can be accessed in the repository. Most useful data in this regard would likely be the building type, design brief, and overall structural system. Such information can inform and trigger queries to the repository, returning similarity-based matches in terms of structure, topology, and/or design requirements. Yet, specific features of retrieved cases, such as system components, material properties, operational performance parameters, etc. would potentially be retrieved in a second phase, which will naturally stimulate the use of BIM and CDE environments. This would in turn enhance further interpretation and learning by the design professionals, simultaneous with the implementation of their domain expertise in the decision-making. The proposed framework will also need to support that initial phase and infer design semantics and characteristics from very limited data. Further investigations are therefore needed to identify the efficiency of the proposed system in the very early design stages.

The devised framework can also be of direct relevance in the technical design phases, where many core decisions are already made, yet specific ones still need to be taken. Such environments rely heavily on digital models and tools, which once again reflects the positioning of the suggested framework in a BIM and CDE context. The above mentioned issues pertaining to availability of data in the early stages are generally not present here. This phase of the design process is strongly characterised by an abundance of data, both in terms of types and representations. As the proposed system aims to leverage exactly this multiplicity of data, it should fit in this part of the design and engineering process. As a result, the workflows characteristic to design practice at this stage would be preserved, apart from the additional presence of precise user-centred recommendations coming in through the BIM and CDE tools.

Using tangible performance data to impact decision-making and prevent errors early in the design phase is increasingly important. Leveraging computational approaches to enhance sustainability-oriented practices, and following an evidence-based path will empower knowledge sharing and reuse, and reduce knowledge vaporization and uncertainty in design decision-making.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Aamodt, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, *7*(1), 39-59.

Agrawal, R., Imielinski, T., & Swami, A. (1993, May). Mining association rules between sets of items in large databases. In *SIGMOD 1993*. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207–216). Washington, DC: ACM.

Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H. & Menzel, K. (2011). Mining building performance data for energy-efficient operation. *Advanced Engineering Informatics, 25*, 341–354.

Akin, Ö. (2014). Necessity of Cognitive Modeling in BIM's Future. In: *Building Information Modeling BIM in Current and Future Practice.* New Jersey: Wiley, pp. 17-27.

Aksamija, A. (2012). *BIM-Based Building Performance Analysis: Evaluation and Simulation of Design Decisions.* Washington, DC: Omnipress.

Alexander, C. (1977). *A Pattern Language*. New York, NY: Oxford University Press.

Alter, S. (2004). A work system view of DSS in its fourth decade. *Decision Support Systems*, *38*, 319-327.

Arnott, D. & Pervan, G. (2008). Eight key issues for the decision support system discipline, *Decision Support Systems*, *44*, 657-672.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web, *Scientific American*, 29-37.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Cambridge, UK: Springer.

Bonino, D. & Corno, F. (2008). DogOnt - Ontology Modeling for Intelligent Domotic Environments. In *Lecture Notes in Computer Science: Vol: 5318*. *Proceedings of the International Semantic Web Conference* (pp. 790-803).

British Standards Institute (2013). PAS 1192-2:2013 Specification for information management for the capital/delivery phase of construction projects using building information modelling.

BuildingSMART (2016). BuildingSMART specifications. Retrieved from http://www.buildingsmart-tech.org/specifications.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*, Upper Saddle River, NJ: Prentice Hall.

Capozzoli, A., Piscitelli, M.S., Gorrino, A., Ballarini, I. & Corrado, V. (2017). Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. *Sustainable Cities and Society*, *35*, 191–208.

Cemesova, A., Hopfe, C. J. & Mcleod, R. S. (2015). PassivBIM: Enhancing interoperability between BIM and low energy design software. *Automation in Construction, 57*, 17-32.

Cheng, Z., Zhao, Q., Wang, F., Chen, Z., Jiang, Y. & Li, Y. (2016). Case studies of fault diagnosis and energy saving in buildings using data mining techniques. In *Proceedings of IEEE International Conference on Automation Science and Engineering* (pp. 646-651). Fort Worth, TX: IEEE.

Davenport, T. H. & Prusak, L. (2000). *Working Knowledge: How organizations manage what they know.* Boston, MA: Harvard Business School Press.

de Souza, C. B. & Tucker, S. (2015). Thermal simulation software outputs: a framework to produce meaningful information for design decision-making, *Journal of Building Performance Simulation*, *8*(2), 57-78.

de Souza, C. B. & Tucker, S. (2016). Thermal simulation software outputs: a conceptual data model of information presentation for building design decision-making. *Journal of Building Performance Simulation*, *9*(3), 227-254.

D'Oca, S. & Hong, T. (2014). A data-mining approach to discover patterns of window opening and closing behavior in offices. *Building and Environment*, *82*, 726-739.

D'Oca, S., & Hong, T. (2015). Occupancy schedules learning process through a data mining framework. *Energy and Buildings*, *88*, 395–408.

Dorst, K. & Cross, N. (2001). Creativity in the design process: co-evolution of problem-solution. *Design Studies*, *22*(5), 425-437.

Eastman, C., Teicholz, P., Sacks, R. & Liston, K. (2011). *BIM Handbook - A guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors* (2nd ed.), Wiley.

Elouti, B.H. (2009). Design knowledge recycling using precedent-based analysis and synthesis models. *Design Studies*, *30*, 340-368.

El-Diraby, T., Krijnen, T. & Papagelis, M. (2017). BIM-based collaborative design and socio-technical analytics of green buildings. *Automation in Construction*, *82*, 59–74.

Fan, C., Xiao, F., Madsen, H. & Wang, D. (2015a). Temporal knowledge discovery in big BAS data for building energy management. *Energy and Buildings*, *109*, 75-89.

Fan, C., Xiao, F. & Yan, C. (2015b). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction*, *50*, 81-90.

Fan, C., Xiao, F., Li, Z. & Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, *159*, 296–308.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*, 37-54.

Fournier-Viger P., Wu C.W., Tseng V.S. & Nkambou R. (2012). Mining Sequential Rules Common to Several Sequences with the Window Size Constraint. In *Lecture Notes in Computer Science: Vol. 7310. Advances in Artificial Intelligence* (pp. 299–304). Berlin, Heidelberg: Springer.

Fu, T.C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *17*, 164–181.

Gaitani N., Lehmann C., Santamouris M., Mihalakakou G. & Patargias P. (2010). Using principal component and cluster analysis in the heating evaluation of the school building sector. *Applied Energy*, *87*, 2079 – 2086.

Goldman, G. & Zarzycki, A. (2014). Smart Buildings/ Smart(er) Designers: BIM and the Creative Design Process. In: *Building Information Modeling BIM in Current and Future Practices.* New Jersey: Wiley, pp. 3-16.

Han, J.W., Kamber, M. & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.) Waltham, US: Morgan Kaufmann.

Hand D., Mannila H. & Smyth P. (2001). *Principles of Data Mining.* Cambridge, USA: MIT Press.

Heylighen, A., Martin, M. & Cavallin, H. (2007). Building Stories Revisited: Unlocking the Knowledge Capital of Architectural Practice. *Architectural Engineering and Design Management*, *3*(1), 65-74.

Heylighen, A. & Neuckermans, H. (2000). DYNAMO: A Dynamic Architectural Memory On-line. *Educational Technology & Society, 3*(2), 86-95.

Ilhan, B. & Yaman, H. (2016). Green building assessment tool (GBAT) for integrated BIM-based design decisions. *Automation in Construction*, *70*, 26-37.

Isikdag, U. (2015). *Enhanced Building Information Models: Using IoT Services and Integration Patterns* (1st ed.). Istanbul: Springer.

Jalaei,. F. & Jrade, A. (2014). Integrating Building Information Modeling (BIM) and Energy Analysis Tools with Green Building Certification System to Conceptually Design Sustainable Buildings. *Journal of Information Technology in Construction (ITcon)*, *19*, 494-519.

Jalaei, F., Jrade, A. & Nassiri, M. (2015). Integrating decision support system (DSS) and building information modeling (BIM) to optimize the selection of sustainable building components. *Journal of Information Technology in Construction (ITcon), 20*, 399-420.

Jun, M.A. & Cheng, J.C.P. (2017). Selection of target LEED credits based on project information and climatic factors using data mining techniques. *Advanced Engineering Informatics*, *32*, 224–236.

Kocaturk, T. (2017). Towards An Intelligent Digital Ecosystem - Sustainable Data-Driven Design Futures. In: *Future Challenges for Sustainable Development within the Built Environment*. UK: Wiley-Blackwells, pp. 164-178.

Liu, Y., Huang, Y.C. & Stouffs, R. (2015a). Using a data-driven approach to support the design of energy-efficient buildings. *Journal of Information Technology in Construction (ITcon)*, *20*, 80-96.

Liu, S., Meng, X. & Tam, C. (2015b). Building information modeling based building design optimization for sustainability. *Energy and Buildings*, *105*, 139-153.

Lu, Y., Wu, Z., Chang, R. & Li, Y. (2017). Building Information Modeling (BIM) for green buildings: A critical review and future directions. *Automation in Construction*, *83*, 134-148.

Mantha, B.R.K., Menassa, C.C. & Kamat, V.R. (2015). A taxonomy of data types and data collection methods for building energy monitoring and performance simulation. *Advances in Building Energy Research*, *10*(2), 263-293.

McGlinn, K., Yuce, B., Wicaksono, H., Howell, S., & Rezgui, Y. (2017). Usability evaluation of a web-based tool for supporting holistic building energy management, *Automation in Construction*, *84*, 154–165.

Miller, C., Nagy, Z. & Schlueter, A. (2015). Automated daily pattern filtering of measured building performance data. *Automation in Construction*, *49*, 1-17.

Mirakhorli, M., Chen, H. & Kazman, R. (2015). Mining Big Data for Detecting, Extracting and Recommending Architectural Design Concepts. In *Proceedings of the 1st IEEE/ACM International Workshop on Big Data Software Engineering* (pp. 15-18). Florence: IEEE.

Park, H.S., Lee, M., Kang, H., Hong, T. & Jeong, J. (2016). Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. *Applied Energy*, *173*, 225-237.

Pauwels, P., Zhang, S. & Lee, Y.C. (2017a). Semantic web technologies in AEC industry: A literature overview. *Automation in Construction*, *73*, 145-165.

Pauwels, P., Krijnen, T., Terkaj, W., & Beetz, J. (2017b). Enhancing the ifcOWL ontology with an alternative representation for geometric data. *Automation in Construction*, *80*, 77-94.

Pearson, J.M., & Shim, J.P. (1995). An empirical investigation into DSS structure and environments. *Decision Support Systems*, *13*, 141-158.

Pena, M., Biscarri, F., Guerrero, J.I., Monedero, I. & León, C. (2016). Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. *Expert Systems With Applications*, *56*, 242–255.

Peng, Y., Lina, J.R., Zhang, J.P. & Hu, Z.Z. (2017). A hybrid data mining approach on BIM-based building operation and maintenance. *Building and Environment*, *126*, 483–495.

Perzylo, A., Somani, N., Rickert, M., & Knoll, A. (2015). An ontology for CAD data and geometric constraints as a link between product models and semantic robot task descriptions. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4197-4203). Hamburg: IEEE.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018a). In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data. In *Proceedings of the 35th CIB W78 Conference* (in press).

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018b). From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches. In Karlshøj & Scherer (Eds.)*, ECPPM 2018. Proceedings of the 12th European Conference on Product & Process Modelling. eWork and eBusiness in Architecture, Engineering and Construction* (pp. 391-399). Copenhagen: CRC Press.

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine*, *11*(5), 68–70.

Polanyi, M.(1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago, IL: The University of Chicago Press.

Rasmussen, M.H., Pauwels, P., Hviid, C.A. & Karlshøj, J. (2017). Proposing a central AEC ontology that allows for domain specific extensions. In *Proceedings of the Joint Conference on Computing in Construction (JC3)* (pp. 237-244).

Richter, K., Heylighen, A., & Donath, D. (2007). Looking back to the future-an updated case base of case-based design tools for architecture. In *Proceedings of the 5th eCAADe Conference* (pp. 285-292).

Ristoski, P. & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, *36*, 1–22.

Sabri, Q.U., Bayer, J., Ayzenshtadt, V., Bukhari, S.S., Althoff, K.D. & Dengel, A. (2017). Semantic Pattern-based Retrieval of Architectural Floor Plans with Case-based and Graph-based Searching Techniques and their Evaluation and Visualization. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods* (pp. 50-60).

Sacks, R., Eastman, C., Lee, G. & Teicholz, P. (2018). *BIM Handbook: A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers* (3rd ed.). Hoboken, NJ: Wiley.

Schlueter, A. & Thesseling, F. (2009). Building information model based energy/exergy performance assessment in early design stages. *Automation in Construction, 182*, 153-163.

Senciuc, A., Pluchinotta, I. & Rajeb, S. B. (2015). *Collective Intelligence Support Protocol: A Systemic Approach for Collaborative Architectural Design.* Mallorca: Springer.

Shadram, F., Johansson, T.D., Lu, W., Schade, J. & Olofsson, T. (2016). An integrated BIM-based framework for minimizing embodied energy during building design. *Energy and Buildings*, *128*, 592-604.

Shen, L., Yan, H., Fan, H., Wu, Y. & Zhang, Y. (2017). An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. *Building and Environment*, *124*, 388-401.

Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., & Carlsson, C. (2002). Past, present and future of decision support technology. *Decision Support Systems*, *33*, 111-126.

Smuts, J.C. (1926). *Holism and evolution*. London, UK: Macmillan.

Soibelman, L. & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, *16*(1), 39–48.

Sonetti, G., Naboni, E. & Brown, M. (2018). Exploring the Potentials of ICT Tools for Human-Centric Regenerative Design. *Sustainability*, *10*.

Starkey, C. & Garvin, C. (2013). Knowledge from data in the built environment. *New York Academy of Sciences 2013 Annals, 1295*(1)*,* 1-9. *The implications of a data-driven built environment.* New York, NY: New York Academy of Sciences.

Strobbe, T., Eloy, S., Pauwels, P., Verstraeten, R., De Meyer, R., & Van Campenhout, J. (2016). A graph-theoretic implementation of the Rabo-de-Bacalhau

transformation grammar. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *30*, 138 - 158.

Thierauf, R.J. (1999). *Knowledge management systems for business* (1st ed.). Westport, CT: Greenwood Publishing Group.

Timmermans, H. (2016). Design & Decision Support Systems in Architecture and Urban Planning. *Proceedings of the 13th International Conference on Design & Decision Support Systems in Architecture and Urban Planning*.

Tucker, S. & de Souza, C.B. (2016). Thermal simulation outputs: exploring the concept of patterns in design decision-making, *Journal of Building Performance Simulation*, *9*(1), 30-49.

Underwood, J. & Isikdag, U. (2010). *Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies* (1st ed.). Hershey, PA: Information Science Reference- IGI Publishing.

United Nations (2010). Sustainable Development. [Online] Available at: http://www.un.org/en/ga/president/65/issues/sustdev.shtml

Vahdatpour, A., Amini, N. & Sarrafzadeh, M. (2009). Towards unsupervised activity discovery using multi-dimensional motif detection in time series. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1261-1266).

Wang, Z. & Srinivasan, R.S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, *75*, 796–808.

Wong, J.K.W. & Zhou, J. (2015). Enhancing environmental sustainability over building life cycles through green BIM: a review. *Automation in Construction*, *57*, 156–165.

Wu S. & Clements-Croome D. (2007). Understanding the indoor environment through mining sensory data—A case study. *Energy and Buildings*, *39*, 1183 – 1191.

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, *75*, 109–118.

Xiao, X., Skitmore, M. & Hu, X. (2017). Case-based reasoning and text mining for green building decision making. *Energy Procedia*, *111*, 417 – 425.

Yalcinkaya, M. & Singh, V. (2015). Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis. *Automation in Construction, 59*, 68-80.

Yarmohammadi, S., Pourabolghasem, R., Shirazi, A. & Ashuri, B. (2016). A sequential pattern mining approach to extract information from BIM design log files. In *Proceedings of the 33rd International Symposium on Automation and Robotics in Construction* (pp. 174-181).

Yu, Z., Fung, B. & Haghighat, F. (2013). Extracting knowledge from building-related data - A data mining framework. *Building Simulation*, *6*(2), 207-222.

Zanni, M.A., Soetanto, R. & Ruikar, K. (2017). Towards a BIM-enabled sustainable building design process: roles, responsibilities, and requirements. *Architectural Engineering and Design Management*, *13*(2), 101-129.

Zhang, C., Beetz, J., & de Vries, B. (2018). BimSPARQL: Domain-specific functional SPARQL extensions for querying RDF building data. *Semantic Web*, *9*(6), 829-855.

Zhao, J., Lasternas, B., Lam, K.P., Yun, R. & Loftness, V. (2014). Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. *Energy and Buildings*, *82*, 341-355.

Zhao, H. & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*, 3586–3592.

Zhou, Q., Zhou, H., Zhu, Y. & Li, T. (2015). Data-driven solutions for building environmental impact assessment. In *Proceedings of the 9th International Conference on Semantic Computing* (pp. 316-319)

Appendix B. Paper II

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data. In: I. Mutis & T. Hartmann (Eds.) *Advances in Informatics and Computing Computing in Civil and Construction Engineering*, pp.19-27, Springer.

https://doi.org/10.1007/978-3-030-00220-6_3

In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data

Ekaterina Petrova[1][0000-0002-8651-0671] , Pieter Pauwels[2][0000-0001-8020-4609], Kjeld Svidt[1][0000-0002-5078-6270] and Rasmus Lund Jensen[1][0000-0001-7008-2601]

[1] Aalborg University, Dept. of Civil Engineering, Thomas Manns Vej 23, Aalborg, Denmark
ep@civil.aau.dk - ks@civil.aau.dk - rlj@civil.aau.dk
[2] Ghent University, Dept. of Architecture and Urban Planning,
J. Plateaustraat 22, Ghent, Belgium
pipauwel.pauwels@ugent.be

Abstract. Cross-domain analytical techniques have made the prediction of outcomes in building design more accurate. Yet, many decisions are based on rules of thumb and previous experiences, and not on documented evidence. That results in inaccurate predictions and a difference between predicted and actual building performance. This article aims to reduce the occurrence of such errors using a combination of data mining and semantic modelling techniques, by deploying these technologies in a use case, for which sensor data is collected. The results present a semantic building data graph enriched with discovered motifs and association rules in observed properties. We conclude that the combination of semantic modelling and data mining techniques can contribute to creating a repository of building data for design decision support.

Keywords: BIM, Semantics, Data Mining, Pattern Recognition, Knowledge Discovery

1 Introduction

Cross-domain analytical techniques such as Big Data analytics, machine learning, semantic query techniques and inference machines have made the prediction of outcomes in building design possible and much more accurate. Research has shown promising advances within the use of machine learning and data mining techniques for model predictive control, metamodelling for design space exploration, grey box modelling and advanced control strategies related to building energy systems, etc. These approaches carry a powerful potential and can directly influence the decision-making process in the Architecture, Engineering and Construction (AEC) industry by infusing it with an evidence-based character. The latter is of direct relevance for high-performance building design, which employs strict performance criteria. Responding to these criteria ideally requires evidence-based multidisciplinary input. Nevertheless, many decisions are still based on rules of thumb and previous experiences, and not on documented evidence. This leads to inaccurate predictions and assumptions regarding input parameters (e.g. occupancy rate), rare revisiting of

analytical and building models during operation, no modification of design assumptions based on actual performance and thus a difference between predicted and measured performance.

If knowledge discovered in building operation would be accessible, a design professional should be able to match the ongoing design with meaningful performance patterns. This article aims to investigate how data from buildings in operation can enable knowledge discovery and provide patterns that can be useful to inform future design processes. In particular, we consider available operational building data related to indoor space use, thermal performance and indoor climate collected from a culture and sports center. This use case is particularly interesting, as the building hosts different spaces such as conference and exhibition halls, ice hockey arenas, training facilities, swimming and wellness facilities, etc. The case provides operational building data captured through a sensor network and existing CAD drawings. From the collected datasets, we distil patterns and represent these so that they can be reusable by deploying the latest technological advances within Knowledge Discovery in Databases (KDD) [1] and semantic data modelling. The considered techniques are not often easily combined, especially not to inform future design decisions, which is the fundamental purpose of this study.

In this article, we first look into the diverse existing computational approaches for data analytics and knowledge discovery (Section 2), and semantic representation of building data (Section 3). In Section 4, we indicate how these data can be combined for knowledge discovery. We thereby suggest a system architecture aimed specifically at that purpose. Section 5 presents the use case we relied on for knowledge discovery, including the results obtained from that use case.

2 Data Analytics and Knowledge Discovery in the AEC Industry

The AEC industry nowadays generates large volumes of data associated with all stages of the building life-cycle. However, the traditional analytics can generate informative reports, but fail when it comes to content analysis [2]. As a result, data mining, pattern recognition and KDD have received major attention, as they can provide reliable results and effectively assist in analysis of data and extraction of knowledge. One definition of data mining is "the analysis of large observational datasets to find unsuspected relationships and to summarize the data in novel ways so that data owners can fully understand and make use of the data." [3] Furthermore, Bishop defines pattern recognition as "the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories" [4]. Finally, KDD represents the overall process of knowledge extraction, with knowledge being the end product of the data-driven discovery and data mining being the step in the process which employs specific algorithms to discover patterns in the given data [5]. Fayyad et al. [1] state that the fundamental objective is to discover high-level knowledge in low-level data and define the transformation steps of raw data into actionable

knowledge, i.e. data selection, preprocessing, transformation, mining and interpretation/evaluation of the discovered knowledge.

Widely accepted data mining categories include classification, clustering, association rule mining, regression, summarization and anomaly detection, targeting either predictive (supervised, directed) or descriptive (unsupervised, undirected) analytics [1, 6]. Supervised approaches describe the qualitative or quantitative relationships between the input and output variables and rely on domain expertise and significant amounts of training data. As a result, discovery of novel knowledge is unlikely, due to the predefined inputs and outputs. Unsupervised approaches (e.g. clustering, association rule mining, etc.), however, excel in discovering the intrinsic structure, correlations and associations in data and do not rely on training data, as inputs and outputs are not predefined. While predictive techniques are backward oriented due to their predefined target, descriptive ones are forward oriented (no explicitly defined target) and make it possible to discover interesting patterns and relationships in the data [7].

Within the high-performance and sustainable building design domain, the use of predictive approaches is usually related to prediction of building energy use and demand [8-10]; prediction of building occupancy and occupant behaviour [11, 12]; and fault detection diagnostics [13, 14]. Unsupervised tasks usually complement and target framework development [15-17]; discovery of patterns in occupant behaviour for improvement of operational performance [18]; and extraction of energy use patterns [19, 20]. Of course, KDD applications in the AEC industry span over a much broader area than the main categories defined above. For instance, Jun & Cheng [21] target high-performance with classification models for sustainability certification evaluation and Peng et al. [22] propose the use of BIM-based data mining approaches for improvement of facility management , etc.

These studies all show promising results when it comes to improvement of the building operation and occupant comfort. However, using knowledge discovery in data to support future design decision-making is an area that is not explored in detail. Studies have explored pattern recognition in simulation data and information extraction from BIM design log files [23], data-driven approaches for energy-efficient design by BIM data mining [24], as well as use of data mining for extracting and recommending architectural concepts [25]. Even though these studies demonstrate promising results within the use of KDD for design decision support, they rely on patterns only in design data. The data analysis results coming from existing buildings can rarely be linked to an early stage design, mainly because the data representations do not match. Thus, this study attempts to explore knowledge discovery in operational building data as a means to improve the decision-making in the performance oriented design process.

3 BIM and Semantic Representations of Building Data

The representation of building information nowadays typically happens using a BIM model, most commonly exchanged using the Industry Foundation Classes (IFC) data

model, which captures building geometry, object properties, as well as semantics. The IFC schema is represented in the EXPRESS information modelling language. Any file exported to IFC is then typically an IFC STEP Physical File (IFC-SPF). Alternative formats for the IFC data model are available in XML, RDF and JSON. In all cases, however, the data model itself is derived directly from the EXPRESS or IFC-SPF format, making it the absolute reference.

Recent research and development initiatives have showed promising results using graph-based data modelling techniques, which are more common in a web environment (e.g. Neo4J, GraphDB). Such approaches are the preferred solution especially when a link needs to be made to outside data that is not typically captured in an EXPRESS-based format (e.g. sensor data, geospatial data). Typically, graph-based approaches focus entirely on the semantics and less on other specific data, such as geometry, large amounts of tabular data, etc. In such case, the semantic graph contains a direct link to the relevant information, which is kept in its original format. Both practice and research thus suggests the use of a graph-based format to capture building data, nevertheless keeping numeric data explicitly out of the semantic graph for computational performance reasons.

Representing semantic building data in a graph format can be done with the available ontologies by the W3C Linked Building Data (LBD) Community Group[1]. This includes a Building Topology Ontology (BOT) [26], a PRODUCT ontology, a PROPS ontology (properties), and an Ontology for Property Management (OPM). Using linked data technologies, links can then be maintained with other data [27], including operational data. For instance, device data can be captured using SAREF[2], and sensor data can be represented using SSN[3] and/or SOSA[4]. For the building performance data, these ontologies do not serve well in case all operational data are targeted. In such case, a tabular format is still a lot more effective. The mentioned semantic ontologies can be used to capture static characteristics, such as averages, min-max values, features of interest, devices, and so forth.

4 Combining Semantics and KDD to Enhance High-Performance Design: Proposed System Architecture

In this article, we consider the combination of KDD (Section 2) and building semantics (Section 3) for the purpose of design decision support. Most importantly, design decision support tools need to re-use the knowledge discovered in the available data through KDD and semantic data modelling. In this section, we focus entirely on discovering patterns using KDD and semantic data modelling, so that a repository of queryable design patterns can be built. Considering that the available data originate

[1] https://www.w3.org/community/lbd/
[2] https://w3id.org/saref
[3] https://www.w3.org/TR/vocab-ssn/
[4] https://www.w3.org/ns/sosa/

from multiple heterogeneous sources, a decentralized structure is preferred, which is most commonly realized using graph database approaches. Using these technologies, one can construct a web of semantic information in a decentralized manner, thereby allowing links between datasets, while respecting their original data structures. Transforming all data to a semantic format is possible and allows direct queries and applying semantic data mining techniques [28]. However, this approach may disallow many highly efficient data mining algorithms that can be used for retrieving useful knowledge. Instead, we propose to store the different kinds of data separately, thereby distinguishing between semantic data, geometric data and operational data (Fig. 1).

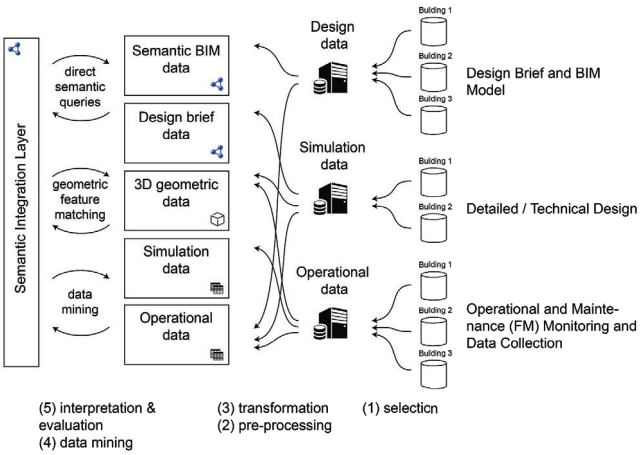

Fig. 1. Proposed system architecture for the combination of semantics and KDD.

We additionally suggest a semantic data integration layer for linking the semantic data model of a building with its numeric representations and dynamic performance parameters. This layer serves as a reference model for the semantics of the different data sources and makes integration possible by pointing from within the semantic graph to web server addresses for operational data streams and geometric data files. As a result, systems accessing this data can recognize the relevant associations.

5 Use Case: Gigantium Cultural and Sports Center

Gigantium is a large cultural and sports center in Aalborg, Denmark, which opened to the public in 1999. Initially, it housed a hall with indoor football and handball courts, a sports hall and meeting facilities. In 2007, two ice skating halls were added, followed by swimming facilities in 2011. Today, Gigantium hosts an ice skating arena and training facility, sports halls, a concert and exhibition hall, swimming and wellness facilities, athletics hall, meeting rooms, a conference room, a cafe, and a

lobby. The total area of the center is about 34,000 m^2. The ice skating arena can host 5,000 spectators and the main hall capacity during concerts is 8,500.

Operational building data is being collected through a sensor network consisting of 35 nodes, divided in all spaces [29]. The nodes monitor Temperature (°C), Relative Humidity (%), Air Pressure (hPa), Indoor Air Quality (Total Volatile Organic Compounds ((TVOC), ppb) and CO2 (ppm)), illuminance (lux) and motion. The purpose of the data collection spans from monitoring indoor climate and thermal comfort, to providing information on space use for maintenance of the facilities. Clearly, the diversity of facilities and activities will be reflected in the collected data. For instance, temperature and relative humidity for meeting rooms, ice hockey arenas, and swimming pool will clearly be different. As a result, this use case provides an ideal dataset that can be used to test the proposed knowledge discovery approach in diverse environments within the same building. Most importantly, the discovered patterns can then inform design decisions related to thermal comfort and indoor climate. For example, persisting issues have been experienced with overheating in the conference room, which has led to a decision to renovate the mechanical ventilation system. The discovered insights would be invaluable to the decision-making related to the system design, by preventing uninformed decisions or use of design parameters that previously led to these issues.

5.1 Capturing the Building Semantics Using a Semantic Graph

As the use case building was built in 1998, there was no BIM model or 3D geometry available as project data. Instead, access was only available to 2D CAD data in PDF format. In this research, we generated a semantic graph from the available data. The spaces are represented using the BOT ontology as *bot:Space* instances. Each of the spaces is linked to its corresponding sensor nodes. These are defined as *bot:Element* and *gig:SensorNode* class instances. The *gig:SensorNode* class is a direct subClass of the *sosa:Platform* class, which is defined by the SOSA ontology to "carry at least one Sensor, Actuator, or sampling device to produce observations, actuations, or samples". Each sensor node hosts sensors, tracking different observable properties (Section 5). The information is described in a graph, following a combination of the BOT and SOSA ontologies, including custom classes and properties (namespace "gig:").

Important to note is that the data values are not directly stored in the semantic graph. Instead, a custom *gig:values* datatype property points to a web address that returns the data values as requested using the HTTP protocol. One is able to add attributes to an HTTP request, thereby setting query parameters such as time frame and refresh rate (e.g. *from=now-30d&to=now&refresh=30*s). The result includes the pointer to the data stream for a *sosa:Result* of a *sosa:Observation*. A full data sample is available[5], yet, access to the sensor data streams is obviously restricted.

[5] http://users.ugent.be/~pipauwel/CIBW87_additionaldata.html

```
inst:room_1
        rdf:type bot:Space ;
        rdfs:label "Main hall" ;
        bot:hasSpace inst:room_2 ;
        gig:hasSensorNode inst:sensorNode_00000097, inst:sensorNode_000000B0,
                inst:sensorNode_00000077 ;
        geom:hasGeometry “2000, 3000, 4000, 6000”^^wkt:linestring.

inst:sensorNode_00000097 rdf:type gig:SensorNode, bot:Element ;
        rdfs:label "00000097" ;
        gig:observation "Space use" ;
        sosa:hosts inst:sensor_00000097_1 ;
        gig:placement "Placed in the middle of the hall, 8m above the floor. " .

inst:result_1 rdf:type sosa:Result ;
        rdfs:label "Result of observation of Relative Humidity" ;
        gig:values "https://gigantium.dk/Gigantium2018instances?orgId=1&datastream=true"
```

Although not in direct focus for this paper, geometry of spaces is also stored in this semantic graph (*geom:hasGeometry*). This representation relies on a Well-Known Text (WKT) and can be used for simple visualization of the relevant spaces in a web-based floor plan layout visualization.

5.2 Knowledge Discovery in Operational Building Data

According to Fan et al. [30], operational building data is essentially multivariate time series data, where each observation is a vector of multiple measurements, and time intervals between subsequent observations are fixed. In that case, knowledge discovery can help capture relationships between variables over particular time periods (frequent repetitive patterns (motifs) and association rules [31]). This article demonstrates the implementation of these approaches on the diverse data streams from the cafe in the lobby. The location is chosen for its varying number of visitors both on a daily basis and during events, thereby minimising the likelihood of discovery of patterns due to regularly scheduled events. The data is collected in the period 12.03-16.05.2018, which constitutes the full available dataset so far. The hourly observations are exported as CSV files and preprocessed to enable motif discovery. Missing data fields are treated with five iterations of multiple imputation by running the Expectation Maximisation bootstrap algorithm in R. Symbolic Approximate Aggregation (SAX) [32] is further applied for dimensionality reduction and transformation of the input time series into strings. The univariate motifs in the multivariate time series data are discovered by identifying Longest Repeated Substrings with Suffix Tree implementation [33]. All repeated instances in the symbolic representation of the time series were identified, as for this effort only disjoint and non-overlapping motifs were considered. Figure 2 shows a graphical representation of the labelled discovered motifs (M1, M2,..., M14) in the sequence of the six variables. Overlapping motifs, as well as motifs contained within other motifs were excluded from observation.

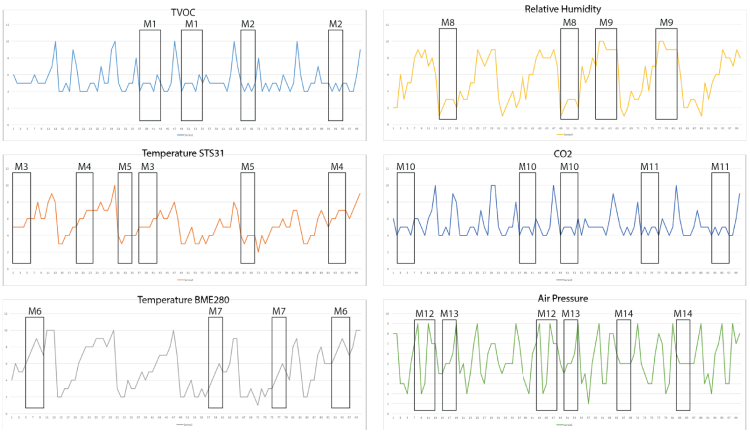

Fig. 2. Discovered univariate motifs (M1-M14) in the observed variables

To enable association rule mining, the discovered motifs are further used to construct a co-occurrence matrix. The columns of the matrix correspond to the motif number and the values for each row (1 or 0) indicate whether an univariate motif occurs or not. For example, M3 co-occurs with M10 and M6. Using the co-occurrence matrix, we obtained 10 sets of co-occurring items for the considered space. Associations between the items of these 10 sets have then been identified by using the association rule mining algorithm defined in [34]. Setting the minimum support and confidence as 0.2 and 0.8 respectively, this results in 13 association rules with support equal to 0.2 and confidence 1. Nine association rules are related to the co-occurrence of M7, M9 and M14. Other association rules are M1 => M10, M3 => M10, M12 => M10, M13 => M8, the last of them being a bidirectional association rule. This means that, for instance, when M12 occurs, the probability of M10 co-occurring is 100%. In this case, the rule indicates an association between observation patterns related to air pressure and CO2. Naturally, the meaning of the discovered rules needs to be interpreted relatively to the design purpose. To be able to use the discovered knowledge, it also has to be connected to the semantic graph in Section 5.1. This can be done by representing the rules in a semantic graph, and linking this graph to the representation of sensor node 00000014, to create a single motif-enriched graph.

6 Conclusion

Knowledge discovered in operational data can be linked directly to a semantic representation of the building and can also be used for retrieving and re-using patterns. In this work, we aimed at making high-performance design rely more explicitly on tangible evidence from operational building data. In order to untap as

much knowledge as possible from available sources, data mining and semantic data modelling are used. The combination of these techniques is not often intensively deployed in an AEC context. Yet, this combination provides great advantages, as formal semantic query can be combined with flexible and high-performing pattern recognition techniques. In this paper, we employ these techniques for the Gigantium Cultural and Sports Center in Aalborg. We hereby relied on the W3C ontologies for linked building data to model the building in direct connection to the available data streams. Furthermore, motif discovery and association rule mining were applied to the sensor data, thereby providing hidden knowledge through the semantic graph. This technique can in future work be used to build a repository that can inform any building designer of high-performing building design techniques.

Acknowledgments

The authors would like to thank Dr. Mads Lauridsen and Aalborg Municipality for providing access to the sensor data used to perform the experiment.

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 37-54 (1996).
2. Soibelman, L., Kim, H.:Data preparation process for construction knowledge generation through knowledge discovery in databases. Journal of Computing in Civil Engineering, 16(1), 39–48 (2002).
3. Hand D., Mannila H., Smyth P.: Principles of Data Mining. MIT Press, Cambridge, (2001).
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, NY, (2006).
5. Piatetsky-Shapiro, G.: Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine, 11(5), 68–70 (1991).
6. Han, J.W., Kamber, M., Pei, J.: Data mining concepts and techniques. 3rd edn. Morgan Kaufmann, Waltham, US (2012).
7. Fan, C., Xiao, F., Li, Z., Wang, J.: Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings, 159, 296–308 (2018).
8. Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H., Menzel, K.: Mining building performance data for energy-efficient operation. Advanced Engineering Informatics, 25, 341–354 (2011).
9. Wang, Z., Srinivasan, R.S.: A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renewable and Sustainable Energy Reviews, 75, 796–808 (2017).
10. Zhao, H., Magoulès, F: A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16, 3586–3592 (2012).
11. D'Oca, S., Hong, T.: A data-mining approach to discover patterns of window opening and closing behavior in offices. Building and Environment, 82., 726-739 (2014).
12. Zhao, J., Lasternas, B., Lam, K.P., Yun, R., Loftness, V.: Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. Energy and Buildings, 82, 341-355 (2014).

13. Cheng, Z., Zhao, Q., Wang, F., Chen, Z., Jiang, Y., Li, Y.: Case studies of fault diagnosis and energy saving in buildings using data mining techniques. IEEE International Conference on Automation Science and Engineering., pp. 646-651 (2016).
14. Pena, M., Biscarri, F., Guerrero, J.I., Monedero, I., León, C.: Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. Expert Systems With Applications, 56., 242–255 (2016).
15. D'Oca, S., Hong, T.: Occupancy schedules learning process through a data mining framework. Energy and Buildings, 88, 395–408 (2015).
16. Fan, C., Xiao, F., Yan, C.: A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automation in Construction, 50, 81-90 (2015).
17. Yu, Z., Fung, B., Haghighat, F.: Extracting knowledge from building-related data-A data mining framework. Building Simulation, 6(2), 207-222 (2013).
18. Xiao, F., Fan, C.: Data mining in building automation system for improving building operational performance. Energy and Buildings, 75, 109–118 (2014).
19. Miller, C., Nagy, Z., Schlueter, A.: Automated daily pattern filtering of measured building performance data. Automation in Construction, 49, 1-17 (2015).
20. Wu, S., Clements-Croome, D: Understanding the indoor environment through mining sensory data—A case study. Energy and Buildings, 39, 1183 − 1191 (2007).
21. Jun, M.A., Cheng, J.C.P.: Selection of target LEED credits based on project information and climatic factors using data mining techniques. Advanced Engineering Informatics, 32, 224–236 (2017).
22. Peng, Y., Lina, J.R., Zhang, J.P., Hu, Z.Z.: A hybrid data mining approach on BIM-based building operation and maintenance. Building and Environment , 126, 483–495 (2017).
23. Yarmohammadi, S., Pourabolghasem, R., Shirazi, A., Ashuri, B.: A sequential pattern mining approach to extract information from BIM design log files. 33rd International Symposium on Automation and Robotics in Construction., pp. 174-181 (2016).
24. Liu, Y., Huang, Y.C., Stouffs, R.:. Using a data-driven approach to support the design of energy-efficient buildings. ITCon, 20, 80-96 (2015).
25. Mirakhorli, M., Chen, H., Kazman, R.: Mining Big Data for Detecting, Extracting and Recommending Architectural Design Concepts. IEEE/ACM 1st International Workshop on Big Data Software Engineering., pp. 15-18 (2015).
26. Rasmussen, M.H., Pauwels, P., Hviid, C.A., Karlshøj, J.: Proposing a central AEC ontology that allows for domain specific extensions. In: Proceedings of the Joint Conference on Computing in Construction (JC3), pp. 237-244 (2017).
27. Pauwels, P., Zhang, S., Lee, Y.C.: Semantic web technologies in AEC industry: A literature overview. Automation in Construction, 73, 145-165 (2017).
28. Ristoski, P., Paulheim, H.: Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics, 36, 1–22 (2016).
29. Rodriguez, I., Lauridsen, M., Vasluianu, G., Poulsen, A.N., Mogensen, P.: The Gigantium Smart City Living Lab: A Multi-Arena LoRa-based Testbed. 15th International Symposium on Wireless Communication Systems, Lisbon, Portugal (2018, in press).
30. Fan, C., Xiao, F., Madsen, H., Wang, D.:Temporal knowledge discovery in big BAS data for building energy management. Energy and Buildings, 109, 75-89 (2015).
31. Fu, T.C.: A review on time series data mining, Engineering Applications of Artificial Intelligence, 17, 164–181 (2011).
32. Patel, P., Keogh, E., Lin, J., Lonardi, S.: Mining Motifs in Massive Time Series Databases. In proceedings of the 2002 IEEE International Conference on Data Mining. (2002).
33. Weiner, P.: Linear pattern matching algorithms, 14th Annual IEEE Symposium on Switching and Automata Theory, pp. 1–11 (1973).
34. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 8, (2004).

Appendix C. Paper III

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019, under review). Data mining and semantics for decision support in sustainable BIM-based design. Submitted to *Advanced Engineering Informatics*.

Data mining and semantics for decision support in sustainable BIM-based design

Ekaterina Petrova[a,*], Pieter Pauwels[b], Kjeld Svidt[a], Rasmus Lund Jensen[a]

[a]*Aalborg University, Department of Civil Engineering, Aalborg, Denmark*
[b]*Ghent University, Department of Architecture and Urban Planning, Ghent, Belgium*

Abstract

Machine learning and semantic web technologies provide an unprecedented opportunity to discover valuable hidden knowledge in the operation of the existing building stock and document it in a reusable, modular and extensible way. Such novel knowledge holds great potential for improving building operation, indoor environmental quality, occupant comfort and future design decision-making. However, the different nature of these technologies and the vast heterogeneity of data sources (sensor data, geometric data, semantic data, etc.) make data and knowledge difficult to combine and reuse in a holistic way. In order to enhance sustainable design practices and make them evidence-based, an appropriate combination of data analysis techniques, semantic data modelling, and legacy storage systems is needed. Therefore, this article exploits the integrated adoption of these technologies and proposes a system architecture for evidence-based design decision support, which is tested with two use case buildings. For both buildings, motif discovery and association rule mining have been performed on collected sensor data to discover frequent patterns and association rules in indoor environmental quality observations. The discovered motifs and rules are represented by a newly developed pattern ontology and then combined with semantic representations of the buildings, including topology, geospatial and product data. The result is a semantic cloud of building data enriched with performance patterns that can be used by design teams as a knowledge base for information retrieval and decision support. To test the information retrieval, the enriched semantic cloud for the two buildings is added to two large repositories of building data. The user-centred federated semantic queries indicate that information can be successfully retrieved from the knowledge base for decision support in evidence-based and performance-oriented design practices.

Keywords: Knowledge Discovery in Databases, Semantic Data Modelling, Building Information Modelling, Sustainable Design, Association Rule Mining, Design Decision Support

1. Introduction

Recent years have shown a rapid co-evolution of technology, advanced analytical approaches and richness of information in the Architecture, Engineering and Construction (AEC) industry. As a result, it is now possible to discover valuable knowledge in the operation of the existing building stock and make it available for reuse with semantically rigorous means. This technological empowerment is of particular importance to contemporary building design, which builds on intertwined arrays of performance targets aiming to minimise environmental impact and enhance energy efficiency, comfort, well-being, health and productivity for the building occupants. Being a multi-dimensional matter traditionally encompassing environmental, economic and social factors, sustainability in the built environment has also been redefined by technology to enable design innovation at product, process and operational levels [1]. The technological evolution has also made it possible to track the built environment's heartbeat by implementation of Building Monitoring Systems (BMS) and sensor networks. Additionally, the progress in methodological approaches, various predictive mechanisms and powerful computational techniques (e.g. machine learning, semantic query techniques, inference machines, etc.) has enabled the prediction of design outcomes and their use to inform decision-making. Combined with advanced Building Information Modelling (BIM) [2, 3], these technological means constitute the industry's aids to define, create, monitor and continuously boost the performance of the buildings of the future.

However, despite these advancements, the performance gap between predicted and measured building performance is still a persisting problem, attributed to multi-faceted reasons spread over the entire building life-cycle [4]. During design, the mismatch can be attributed to (1) inaccurate predictions and assumptions related to analytical input parameters (e.g. occupant behaviour, HVAC demand etc.), (2) errors in modelling and lack of collaboration, and (3) a lack of feedback loop from operation to design [5]. Advanced technology is used for the creation of BIM models, but the fragmentation of the different stages of the building life-cycle leads to those models being rarely reused or revisited during building operation. Similarly, the implemented design assumptions remain isolated in the design phase and are seldom modified to account for actual performance. That also includes inconsistencies due to influence from dynamic variables related to external conditions, occupant behaviour and

*Aalborg University, Department of Civil Engineering, Thomas Manns Vej 23, 9220 Aalborg, Denmark

Email addresses: ep@civil.aau.dk (Ekaterina Petrova), pipauwel.pauwels@ugent.be (Pieter Pauwels), ks@civil.aau.dk (Kjeld Svidt), rlj@civil.aau.dk (Rasmus Lund Jensen)

changes in operation. Finally, the lack of data integration and cross-domain data sharing additionally contributes to the existence of the performance gap [6].

These issues are further magnified by the systematic use of rules of thumb and previous experiences to support decision-making, instead of relying on evidence for using particular design approaches or parameters. As previously indicated in Petrova et al. [7], project-specific expertise is essential, but hardly transferable between projects and teams. Thus, such expertise remains captured within the boundaries of the individual projects, even if they reflect best practices and can positively influence future designs based on various levels of similarity. Additionally, the decisions typically aim to fulfil the current needs of the design intent relative to the performance of the building at the time of completion. As a result, future needs due to significant changes in conditions (e.g. environmental conditions) are underestimated. And while the previously mentioned richness of data is strongly recognisable in the large datasets generated during design, construction and operation of buildings, these datasets are seldom reused to inform future building design on a holistic level.

Therefore, the main goal of this research effort is to bridge the gap between the experience-driven and evidence-based approach to design, building on knowledge discovered in operational building data and disparate project data repositories. As suggested in Petrova et al. [7], the dynamic interplay between knowledge discovery techniques and semantic data representation methods can serve as the much needed catalyst for enhancing future design decision-making with the evidence-based character that it is lacking. The novelty of the approach proposed in this article is in the hybrid deployment of BIM for reuse of data in the early stages of building design, use of Knowledge Discovery in Databases (KDD) [8] approaches for hidden knowledge discovery in building data, and implementation of semantic data modelling techniques for knowledge representation and retrieval in the design environment. An indication of how these can be implemented in support of the design team is given in Fig. 1.

The article builds on the initial implementations described in Petrova et al. [9, 10] and aims to demonstrate how knowledge discovered in building operation can be transformed into input for a performance-driven design decision support system. As seen in the mentioned studies, knowledge discovered in operational and project data holds significant potential when it comes to informing design decision-making. Therefore, the objective of this article is to further pave the road towards evidence-based sustainable design relying on relevant knowledge obtained from building operation and thereby achieve the necessary holistic view towards the built environment.

The following section summarises the results from the first fundamental building block of the study, namely an extensive literature review in the areas of knowledge discovery and semantic data modelling for building performance improvement. Section 3 presents the adopted methodology and documents key choices in terms of motif discovery in operational data, as well as knowledge representation and retrieval. Section 4 documents the suggested BIM-based design decision support system, thereby indicating how the different types of data are handled and included in an overall system architecture. Sections 5 and 6 discuss the implementation and results obtained for two example buildings. Finally, Section 7 concludes this article and presents considerations for future work.

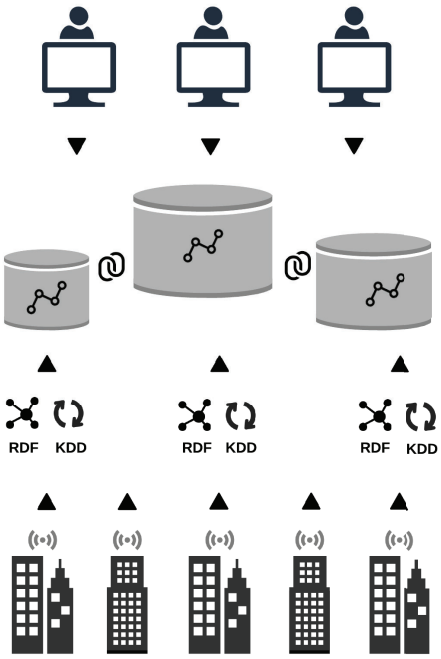

Figure 1: Conceptual overview of the proposed system architecture, which requires data from existing buildings (bottom) to be stored in repositories hosting building information and knowledge discovered in that building data (middle), so that these data can eventually be re-used by diverse end users (top).

2. State of the art

The fundamental topics that the state of the art review encompasses are displayed in the bibliographic timeline in Fig. 2. The purpose of this overview is to provide an insight into the structure and dynamics of the targeted interdisciplinary knowledge domain, including the core areas and the connections between them. Therefore, the following section outlines the state of the art contributions within knowledge discovery and representation, semantic data modelling and sensor data processing technologies, and the way they are applied for building performance improvement and design decision support.

2.1. Knowledge Discovery in Databases

The concept of knowledge discovery in large amounts of data was pioneered by Piatetsky-Shapiro [11] and later further outlined by Fayyad et al. [8]. These seminal works define what is

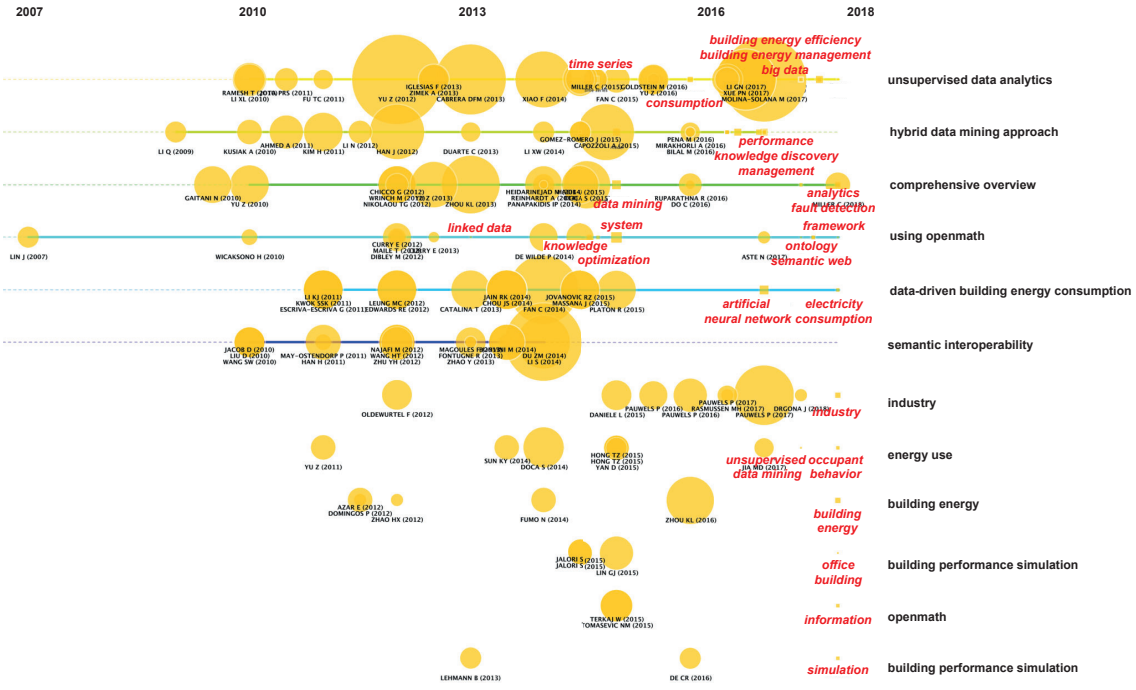

Figure 2: Bibliographic timeline analysis of the reviewed literature according to author information, abstract, keywords and cited references. The horizontal axis represents the timeline, and the vertical axis represents the diverse topical clusters found in the literature and their emerging keywords.

nowadays referred to as KDD, as well as the essential steps to undertake for extraction of high-level knowledge in low-level data, i.e. selection, pre-processing, transformation, data mining and interpretation/evaluation of results. In that context, Hand et al. [12] in turn define data mining as *"the analysis of large observational datasets to find unsuspected relationships and summarise the data in novel ways so that data owners can fully understand and make use of the data"*.

Both research and practice in AEC have to some extent recognised the potential of KDD for discovery of unsuspected hidden patterns and relationships in data, especially because of the inability of traditional analytical approaches to reveal insights in an efficient way. Indeed, data constitutes unstructured facts and figures that in their raw form can hardly impact decision-making, whereas knowledge implies know-how, contextualisation, meaning and understanding. Therefore, the interpretation and contextualisation of data mining results is essential to decision support and performance optimisation in the domain.

2.1.1. KDD according to data source and purpose

Fayyad et al. [8] summarise six main data mining categories, i.e. classification, clustering, association rule mining, regression, summarisation and anomaly detection. Han et al. [13] divide these into two main categories: predictive (supervised) and descriptive (unsupervised). With regards to the input data source, Lausch et al. [14] distinguishes predominantly between numerical and categorical data, text, web, media, time series and spatial data mining. In the AEC industry, spatio-temporal input data is of high importance, considering that a lot of data links a building object in a given location to a particular observation at a given time. In the current context, spatio-temporal data mining can target spatial data from BIM models augmented with time series data from BMS. Fu [15] defines time series data as a collection of observations made chronologically, which are large in size, high in dimensionality and characterised by a necessity of continuous updates. Based on the purpose, various methods for temporal knowledge discovery exist, e.g. events, clusters, itemsets, motifs (frequent sequential patterns), discords (infrequent sequential patterns), anomalies and association rules.

Shekhar et al. [16] rightfully indicate that extracting interesting patterns and associations from such complex and multidimensional data with plenty of dependencies and spatio-temporal correlations is more difficult than mining traditional numeric and categorical data. Machine learning techniques targeting those input data can be of particular value to the construction industry, but the variability in data types and structures further underlines the importance of tailoring the knowledge discovery process and the employed algorithms to the specific data and goals.

2.1.2. *Knowledge discovery for building performance improvement and design decision support*

A significant body of literature explores the use of supervised and unsupervised techniques [17] for the purposes of building performance optimisation, building energy management and efficiency enhancement [18, 19], prediction of energy consumption and energy saving [20, 21, 22, 23, 24], as well as fault detection and diagnosis [25, 26]. Fan et al. [27] demonstrate the potential of temporal knowledge discovery by using energy consumption pattern clustering and association rule mining for improvement of building operation, detecting abnormal system operation and preventing deficit flow. Capozzoli et al. [26] also state that anomalous operation of equipment and building control systems has a large contribution to the performance gap and put a strong focus on the importance of characterising energy consumption patterns over time. As a result, numerous research efforts have explored the benefits of data mining for understanding the behaviour of buildings, predicting future abnormalities in operation and thereby improving building performance [28, 29, 30, 31]. Researchers also investigate the possibilities for improvement of decision making in relation to energy efficiency as a fundamental attribute of building performance and sustainability. Fan et al. [32] use gradual pattern mining to discover co-variations among influential numerical building variables. Fan et al. [33] propose a framework employing interpretable machine learning techniques to help explain and evaluate predictive energy performance models and avoid failure in predictions.

Fan et al. [18] and Miller et al. [34] present extensive reviews of the application of unsupervised analytics for extracting useful insights from operational building data. Miller et al. [34] hereby shed light on the potential of visual analytics as a means to support human interpretation of the analytical results. In another seminal study, Miller et al. [35] address another important issue, namely automating the discovery of behavioural insights in large, unstructured datasets. Other research efforts investigate the potential of dedicated recommendations to the building occupants for reduction of energy consumption [36], discovery of relationships between various building features with significant impact on energy performance [37], the impact of feature engineering on the accuracy of machine learning algorithms for building energy data mining [38], the efficiency and accuracy of different forecasting models for energy consumption prediction [39].

With regards to the use of KDD for bridging the gap between predicted and actual performance, the literature review identified building occupancy as another topic of significant importance [40]. In an comprehensive overview, Zhang et al. [38] underline that understanding occupant behaviour is critical to performance optimisation due to its hardly predictable nature. Critical here are window opening, lighting control and space heating/cooling, as well as methods for data collection and behaviour modelling. D'Oca et al. [41] also state that understanding human behaviour in terms of energy use holds a significant potential when it comes to reducing operating costs, improving indoor environmental quality, etc. As a result, numerous research endeavours target in-depth understanding of occupant behaviour [42, 43, 44, 45].

Other approaches in the performance-oriented design and engineering domain include the use of data mining for development of cost-effective retrofit strategies [46], prediction of cost and schedule performance of green building projects based on early stage variables [47], analysis of the influence of project variables on primary energy demand [48], decision support for definition and achievement of sustainability certification targets [49, 50]. Some of the most recent approaches explore the application of reinforcement learning methods for development of autonomous building energy management systems, as well as performance optimisation by exploiting the latest advancements in sensor technologies and advanced control algorithms [51].

Several research initiatives also attempt to reuse measured performance data and thereby improve the accuracy of design input, simulation and output. For instance, Garrett and New [52] present a methodology for autonomous calibration of building energy models to measured hourly energy usage data. In a similar effort, Tronchin et al. [53] use parametric simulation to increase the robustness of performance estimates in the design phase, while maintaining the fundamental relationship with the operational phase by continuous model calibration based on monitored performance.

In terms of the use of KDD approaches for (design) decision support and building performance optimisation, Peng et al. [54] present an alternative approach to improving building operation by mining BIM data and using the insights to provide recommendations and warnings for maintenance efficiency and improvement of resource use. Jin et al. [55] propose a method for automatic learning of spatial design knowledge from BIM data by the use of clustering and feature extraction. Other research efforts address data-driven design of energy-efficient buildings by mining BIM data [56], extracting 3D modelling patterns from temporal BIM log text data [57] and mining simulation data for energy efficient building design [58]. Using building performance simulation data for the creation of a knowledge base of patterns and its significance to design decision support has also been discussed [59, 60].

As seen in the performed literature study, KDD approaches hold significant and diverse potential when it comes to decision support for building performance optimisation. That potential in itself has been a subject of investigations, aiming to assess the usefulness of KDD to the AEC industry [61, 62, 63]. Ahmed et al. [63] highlight current challenges and drivers for data mining in the industry based on a dedicated workshop with 65 academics and industry professionals. The results point out sustainability and decision support systems as two of six main drivers. The study also shows that the greatest potential for data mining applications lies within design, construction, sustainability and energy analysis, forensic analysis and reuse of digital components. Feedback loop from operation to design is listed as most important when it comes to the design process [63].

The objectives of this research effort align with the KDD literature review results concerning design decision support,

knowledge reuse and feedback loop from building operation to design. And while the potential of KDD seems to be recognised, the AEC industry is still lagging behind on some fundamental implementations, necessary to deploy the full potential of knowledge discovery for decision support in performance-based design. All studies underline the need of human expert interpretation of results, regardless of the level of sophistication of the used algorithms. In addition, despite the recognised need of a feedback loop between building operation and design, such implementations have not been explored. In that context, solutions for evidence-based decision making exist, but they are usually dedicated to the phase of origin of the data. In other words, knowledge discovery in design and simulation data is used for improvement of decision making during the design phase, and mining of measured data is used for improvement of the operational phase, occupant comfort and BMS. Knowledge reuse across the phases of the building life-cycle is rarely explored and usually remains on conceptual level.

2.2. Semantic data modelling

Further to the advancements in KDD, a lot of progress has been made in the formalisation of knowledge and meaning: semantics. Most of this progress has taken place in the context of the web. Even though the web is used for multi-faceted information exchanges, key evolutions focus on the representation of semantics.

2.2.1. Web of Data

From a web of documents, the World Wide Web has now evolved into a 'Web of Data' (Linked Open Data cloud)[1] [64]. The Web of Data relies on a triple subject - predicate - object structure to compose directed labelled graphs (Fig. 3), which together form a web of semantically interlinked datasets.

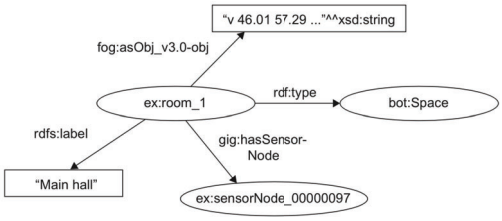

Figure 3: Subject - predicate - object structure for all information in the semantic web.

The term Linked Data was coined by Tim Berners-Lee in 2006[2], where the four rules of linked data were laid out, namely: *"(1) Use URIs as names for things; (2) Use HTTP URIs so that people can look up those names; (3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL); (4) Include links to other URIs, so that they can discover more things."* These rules are the basis of the push towards publishing 5-star open data[3], which implies defining data according to the Resource Description Framework (RDF)[4] data model and interlinking it with other RDF-based datasets available on the web, which constitute the LOD cloud.

The Web of Data relies on vocabularies (ontologies) so that data is typed and can easily be used in combination with query and rule languages such as SPARQL. These ontologies can be defined using RDFS and OWL[5]. They give meaning or semantics to the data, constituting the Semantic Web as it was conceived as early as 2001 by sir Tim Berners-Lee [65]. This semantic network is defined as *"an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."*.

As indicated by Lausch et al. [14], a lot of data can be mined from this data source as well, even though this requires a different approach compared to traditional data mining. In that sense, the mining mostly requires devising intelligent semantic queries (e.g. SPARQL).

2.2.2. Linked Building Data

Because of their potential, linked data and semantic web technologies have received major attention in the AEC industry in the past decade. A comprehensive overview on this topic was performed by Pauwels et al. [66]. Among the most notable initiatives is the early work in transforming the Industry Foundation Classes (IFC) into an OWL ontology (ifcOWL) [67, 68]. This work resulted in the creation of the BuildingSMART Linked Data Working Group (LDWG[6]) and the W3C Linked Building Data Community Group (W3C LBD CG[7]), which aim at standardising the representation of building data over the web.

The ifcOWL ontology is designed according to three main criteria [68], one of which states that *"the ifcOWL ontology should match the original EXPRESS schema as closely as possible"*, even allowing a round-trip conversion process (lossless conversion). However, this has resulted in a very big ontology, which resembles the IFC schema almost completely, i.e. difficult to extend, complex, and not modular. Therefore, several other initiatives aim at defining an ecosystem of smaller, modular and extensible domain ontologies for Linked Building Data [69] (Fig. 4). The LBD concept revolves around a small central Building Topology Ontology (BOT) [70], from which alignments can be made with other domain ontologies [71], such as SAREF [8], DogOnt [72], building product ontologies [73], and so on. Another set of ontologies focuses entirely on 3D geometric data, which is typically a lot harder to represent with linked data approaches [74]. Such geometry constitutes another separate module in the realm of linked building data. How various kinds of geometry may be linked

[1] http://lod-cloud.net/state/
[2] http://www.w3.org/DesignIssues/LinkedData.html
[3] http://5stardata.info/
[4] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
[5] http://www.w3.org/TR/2012/REC-owl2-primer-20121211/
[6] https://technical.buildingsmart.org/community/linked-data-working-group/
[7] https://www.w3.org/community/lbd/
[8] http://ontology.tno.nl/saref/

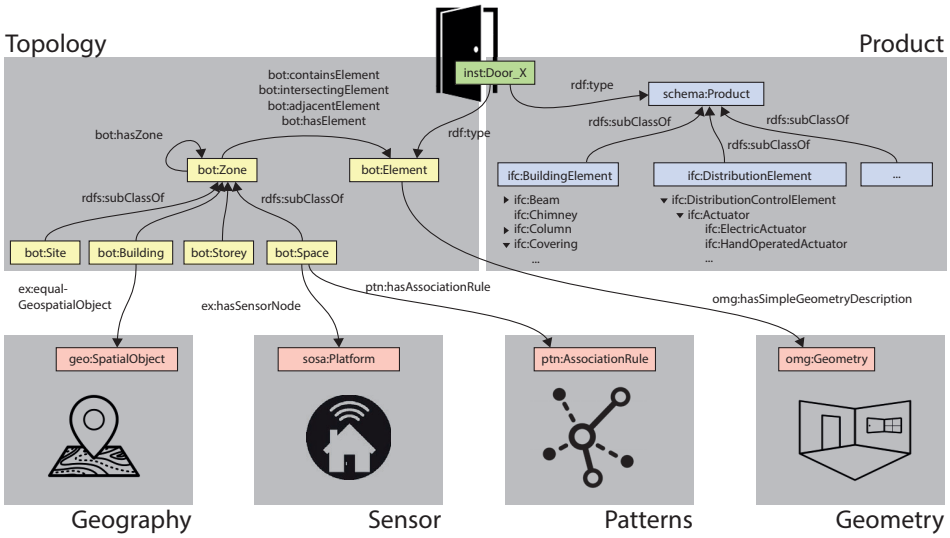

Figure 4: Conceptual overview of the modules and ontologies in the linked building data cloud, based on the work in the W3C LBD CG.

to building data and geospatial data can be found in McGlinn et al. [75].

2.2.3. *Semantic sensor data*

Of particular importance for the current work is the module pertaining to sensor data in the set of LBD ontologies. A key issue related to the representation of sensor data, is the heterogeneity of sensor data sources and environments [76]. Monitored data is usually represented in different ways depending on the sensor network and devices used. The data models and schemas differ just as much. That leads to several compatibility, interoperability and representation issues. To tackle those, research efforts propose various solutions such as semantic annotation of sensor data [77], providing ontology-based access to data [78], using SPARQL queries with streaming extensions to access observations [76], etc. A broader overview of semantic sensor net ontologies, mapping and querying is given in Wang et al. [79]. Most of these works aim at reformatting sensor data so that it is accessible through a semantic query interface. This requires mapping, annotating, and/or processing the sensor data into an alternative format.

Figure 5 shows some of the most often used means to store sensor data and make them accessible. Sensor data is harvested by devices (bottom in Fig. 5) and is either stored directly in an SQL store (bottom left), or is immediately processed using stream processing technologies [80]. Such techniques make the raw data available, typically in a direct API interface (top left in Fig. 5). Alternatively, it has been suggested to make the sensor data available as linked data, and integrate it with other semantic data. This process is represented by the triple store (top right in Fig. 5). Two key elements for transferring raw sensor data (either SQL or stream-based) into semantic sensor data are (1) the ontology, and (2) the mapping mechanism (middle right in Fig. 5). A number of mapping mechanisms (e.g. D2RQ, R2RML, etc.) allow to translate the sensor data in semantic sensor data, either in the form of data dumps or as real-time mappings.

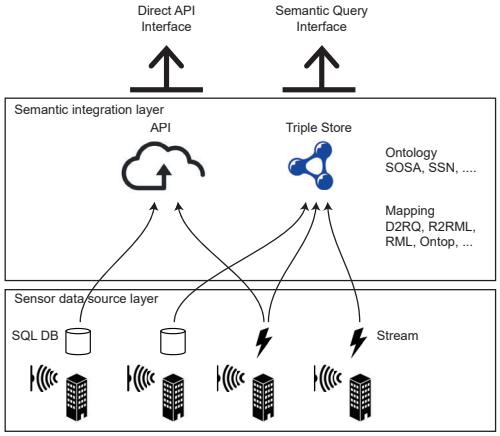

Figure 5: Diagrammatic overview of the ways in which sensor data can be made available to an end user application (inspired by Wang et al. [79]).

Two main ontologies that can be used for the representation of sensor data are SEAS [81] and SSN[9]. A number of recent works have looked into the semantic representation of sensor

[9]https://www.w3.org/TR/2017/CR-vocab-ssn-20170711/

data in combination with a modular LBD approach [82, 10, 69]. A distinct difference can be found in the storage mechanisms for sensor data. For instance, Rasmussen et al. [82] and Schneider et al. [69] store all sensor data in the RDF graph, which results in data representations as shown in Listing 1[10].

```
# SENSOR AND PROPERTY (MODELLED BY ENGINEER)
inst:room_4b80808e-2f04-46a0-b84d-0ad6ee9d6b1b-0012a494
    a bot:Space ;
    bot:containsElement inst:room_04.196-Temp-Sensor .

inst:room_04.196-Temp-Sensor
    a sosa:Sensor , dog:TemperatureSensor ;
    sosa:observes inst:room_04.196-Temp .

inst:room_04.196-Temp
    a sosa:ObservableProperty .

# OBSERVATION (OUTPUT FROM BMS)
inst:room_04.196-Temp-obs0
    a sosa:Observation ;
    sosa:hasFeatureOfInterest inst:room_4b80808e-2f04-46a0-b84d-0
      ad6ee9d6b1b-0012a494 ;
    sosa:hasResult "22.8 Cel"^^cdt:temperature ;
    sosa:madeBySensor inst:room_04.196-Temp-Sensor ;
    sosa:observedProperty inst:room_04.196-Temp ;
    sosa:resultTime "2017-09-16T16:21:54+01:00"^^xsd:dateTime .
```

Listing 1: Sensor data directly embedded in an RDF graph.

Petrova et al. [10] adopts a different approach, in the sense that the sensor data is not fully embedded in the RDF graph. Instead, the sensor data is maintained in its native storage environment, which has a direct API interface (cfr. Fig. 5, left), and a direct reference to that location is embedded in the graph instead. End user applications can then parse the much smaller graph, follow the API links, and fetch sensor data as needed. Such an approach is particularly valuable in cases where real-time data is continuously collected and data analysis does not rely only on datasets of past observations (historical data).

2.2.4. *Semantic approaches to building performance improvement and design decision support*

In terms of the recognised performance gap, this article already indicated that the fragmentation of the industry and the lack of data integration have a large share in its causes [6]. Curry et al. [83] demonstrate the potential of linked data approaches for breaking the isolated information silos and creating a well-connected graph of building data, thereby achieving a holistic perspective on building management. Hu et al. [6] also address the need of cross-domain data sharing in the industry and underline how combining traditionally separate data sources (e.g. linking occupancy patterns to building operation) may enable the discovery of novel performance insights. That includes keeping the different data in the most appropriate format according to its type and sharing it on demand, e.g. linking timeseries data in relational databases with contextual semantic building data [6, 10]. Semantic interoperability in building operation for energy performance optimisation has also been discussed in detail [84]. Corry et al. [5] further expand the contribution to reduction of the performance gap by introducing a performance assessment ontology and framework aiming to transform heterogeneous building data into semantically enriched input for performance analysis. Hu et al. [85] propose an automated performance assessment approach relying on an integration between OpenMath and linked data for evaluation of performance metrics extracted from time series data.

Other research efforts include combining linked data, scenario modelling and complex event processing for building performance optimisation [86], a knowledge-based building energy management system using Artificial Neural Networks, Genetic Algorithms, and Decision Tree rules for building environment optimisation through recommendations [87], automated code compliance checking using BIM data and monitored environmental data from sensor networks [88], an ontology for the standard definition of buildings and related energy efficiency concepts [89], and a smart prediction assistant combining semantic web technologies and KDD for energy efficiency prediction in tertiary buildings [90].

3. Methodology

As seen in the state of the art review, semantic technologies can be effectively used to represent building data, and data mining techniques can help discover hidden knowledge in the performance of the buildings. Yet, no studies have attempted to combine both into a decision support system to provide meaningful input to an end user and establish the missing feedback loop from operation to design. Semantic queries by themselves cannot provide the diversity of insights that can be obtained with data mining techniques. On the other hand, relying solely on data mining cannot provide an integrated view over the diverse datasets or any retrieval opportunities. Additionally, data mining results by themselves lack any semantic expression. Therefore, the diverse data and the discovered knowledge need to be available, semantically enriched and dynamically linked to allow retrieval and design decision support. By doing so, we target a reconciliation between the statistical and symbolic branches of data science (Fig. 6) in a system setup that can enhance human decision-making in sustainable design practice.

The combined use of data mining and semantic web technologies requires to overcome a number of challenges. These two sets of technologies are very different from each other and any system architecture needs to take this into account. From the performed literature review, the following challenges can be summarised:

1. *Multiplicity of data mining algorithms:*
   Data mining algorithms are powerful, abundant and versatile, but selecting an appropriate predictive and/or descriptive mechanism requires high level expertise. Decisions related to data selection, pre-processing, algorithm selection and fitting are in the hands of the analyst.
2. *Manual work in applying data mining methods:*
   Data mining methods usually require a lot of (expensive) manual pre-processing and post-processing tasks, which hinder the creation of a fully automatic system.

[10] from https://github.com/TechnicalBuildingSystems/OpenSmartHomeData

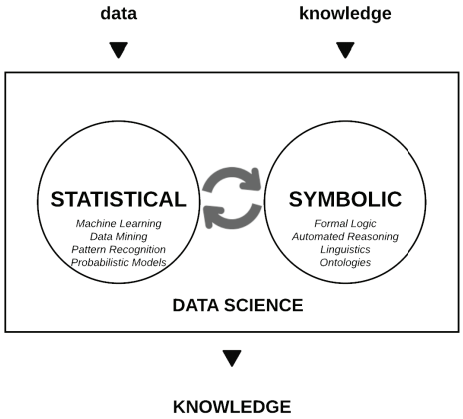

Figure 6: Statistical and symbolic constituents of data science (based on Hoehndorf and Queralt-Rosinach [91]).

3. *Interpretation of data mining results:*
   While data mining algorithms can extract unsuspected patterns and relationships from data in an efficient way, a human expert is still needed for the interpretation of their meaning.
4. *Using semantic data in data mining algorithms:*
   Most data mining methods target traditional datasets, including tabular data, image data, textual data, etc.
5. *Capturing data mining results in semantic graphs:*
   Data mining methods typically result in patterns, which are often directed to a human end user, not a machine.

In the remainder of the article, we look specifically into the discovery of frequent repetitive patterns from time series data and the use of semantic data modelling for representation and storing of the discovered knowledge. Most importantly, that includes combining both into one system aiming to bring back the discovered knowledge to the design team and thereby achieve the targeted feedback loop. Based on the conclusions of the literature review, we devise a system architecture and showcase its implementation with two use cases.

To be able to influence design practice, design teams need to be presented with the discovered knowledge in a way that is meaningful and implementable in their workflows, without disrupting and fragmenting them. That may be in the form of user-centred recommendations or relevant patterns matching the design intent and performance targets. Therefore, in this study, we connect the active design environment to a repository that hosts the discovered performance knowledge from existing buildings.

3.1. Motif discovery in operational building data

To capture relationships between variables in indoor environmental quality data over particular time periods, we perform motif discovery and association rule mining [15]. Missing data fields are treated with five iterations of multiple imputation by applying the Expectation Maximisation bootstrap algorithm. Symbolic Aggregate Approximation (SAX) [92] is then applied for dimensionality reduction and transformation of the time series data into strings. To discover univariate motifs in the multivariate time series, we then identify the Longest Repeated Substrings (LRS) in the SAX strings with a Suffix Tree implementation [93]. In this effort, we consider only disjoint and non-overlapping instances of the frequent patterns. To enable the discovery of association rules, we then use the discovered motifs to compute a co-occurence matrix. Mining of association rules is performed by the use of a frequent-pattern-based method (FP-growth) as defined by Han et al. [94].

3.2. Knowledge representation and retrieval

For knowledge representation and storage, we use the Resource Description Framework (RDF) data model. We choose this method to enrich the discovered patterns in a semantically meaningful way, which is needed for external information retrieval. We hereby primarily use the ontologies proposed by the LBD community group. In order to provide a good basis for decision support, we build a repository of building data by transforming a set of IFC files into RDF graphs using the IFC to LBD conversion software[11]. These RDF graphs are stored in a Stardog[12] repository (knowledge graph platform), which functions as a knowledge base. In addition, the formal RDF representations of the two use case buildings are also added to the knowledge base. These use case buildings are enriched with the observation data and pattern data obtained using the method explained in Section 3.1. A pattern (`ptn:`) ontology is created to be able to represent the data mining results in a semantically meaningful way. Finally, the repository is queried using federated SPARQL queries, showing to what extent the data can be retrieved for evidence-based design decision support.

4. Proposed system architecture

4.1. Data handling according to source

Building data forms the starting point for building a knowledge repository from which knowledge can be retrieved by an end user, in this case a design team working in a BIM environment. The following building data can be accessible in such context:

- Sensor data:
  Many buildings are equipped with BMS and sensor networks, which collect data points and track the performance of the building.
- Textual documents:
  Textual documents pertaining to the particular project may be a valuable source of information and put the discovered knowledge in context, e.g. design brief.

[11] https://github.com/jyrkioraskari/IFCtoLBD
[12] https://www.stardog.com/

- Drawing materials:
  Many existing buildings have a set of building plans associated, which give a two-dimensional indication of the building structure.

- Graphs:
  In a number of cases, more complex semantic building data is available, typically in the form of IFC files that contain element types, materials, and overall building structure in a semantically structured form.

- BIM models:
  For some existing buildings, BIM models are available, or can at least be obtained from laser scans and a scan-to-BIM approach.

Textual documents and drawings are the most difficult and expensive to reuse, because they are typically unstructured, at least from the perspective of a machine. BIM models are very valuable resources, and BIM data is ideally reused in a neutral format, as an open semantic graph (IFC or other format). Considering the work in the area of linked building data, such data can easily and meaningfully be represented in a semantic network. As argued in Pauwels et al. [74] and McGlinn et al. [75], it is considered less useful to include 3D geometric data as a full semantic model in this graph. We rather suggest to link to geometry from within the graph, whereby geometry can be represented in any format, also binary. The graph functions as a central structuring element, hence it is put central in our proposed data storage mechanism (Fig. 7).

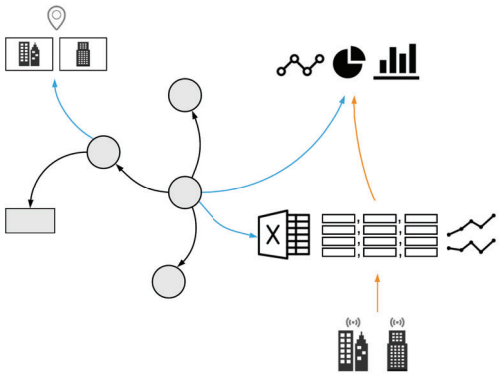

Figure 7: Diagrammatic overview of how data can be handled by the suggested system, with sensor data harvested (bottom right) into comma-separated values (orange bottom right) and mined for patterns (orange top right). Both the mined data and the resulting patterns, as well as other documents (plans, BIM models, geolocation, etc.) can be linked to (blue arrows) from a central semantic graph (left).

Sensor data is key, but even though it can be captured in a semantic network (cfr. SEAS and SOSA/SSN ontologies), this approach is not considered the most efficient, as was indicated in Section 2.2.3. Sensor data have a significant size, especially when data points are stored in continuous streams. Furthermore, the triple structure of the knowledge graphs is not optimal for representation of sequential and ordered data streams [74]. Finally, most of the data mining algorithms rely on traditional formats and storage mechanisms, which are relational databases and/or sets of comma-separated values (see Section 2.1).

Therefore, we suggest to store building data in a semantic RDF graph, ideally combining this graph with raw sensor data through the Application Programming Interfaces (APIs) of their legacy systems, as proposed in Petrova et al. [10]. If no such API or legacy system is available (e.g. only historical data or no direct database or stream access), an RDF-based triple store may be used to represent the sensor data, thereby following the approach suggested by Schneider et al. [69] and Rasmussen et al. [82]. Yet, the semantic graph forms the backbone of the repository. As web technologies and principles are used, this repository architecture can be replicated in many places globally, and linked together to form a web of linked building data enriched with sensor data and performance patterns.

4.2. Representation and storing of data mining results

As already indicated, the power of sensor data lies in the patterns and association rules that can be discovered within them. Hence, to be able to use them in a decision support system, they need to be machine readable and reusable. This can be achieved by storing the discovered knowledge in an enhanced linked building data graph. By storing raw data in their native structures (e.g. storing sensor data in SQL stores and data streams), they are more amenable to be used by data mining algorithms. Hence, such native stores are preferred over RDF-based semantic graphs in this case.

As data access happens through either a direct API access or a semantic query interface (see Fig. 5), it is important that the results of the data mining algorithms are also available through these interfaces. However, data mining results usually require human interpretation. Thus, the best option is to store the raw data mining output in the information system, and make its visual representations available to an end user for interpretation.

4.3. System architecture

Data and algorithms need to be combined in a useful manner, responding to an appropriate web-based system architecture. Such an architecture is proposed in Fig. 8. This system architecture shows how we aim to combine applications (top application layer), including active design environments (BIM tools, parametric design tools, etc.) with a solid set of information repositories.

As suggested in Petrova et al. [7], such an information architecture allows the integration of heterogeneous data sources, enables federated query techniques over diverse data repositories for advanced information retrieval, and provides a well-defined data structure to capture building semantics. Such infrastructure is furthermore entirely compatible with data mining algorithms that function with sensor data represented in legacy systems (bottom in Fig. 8).

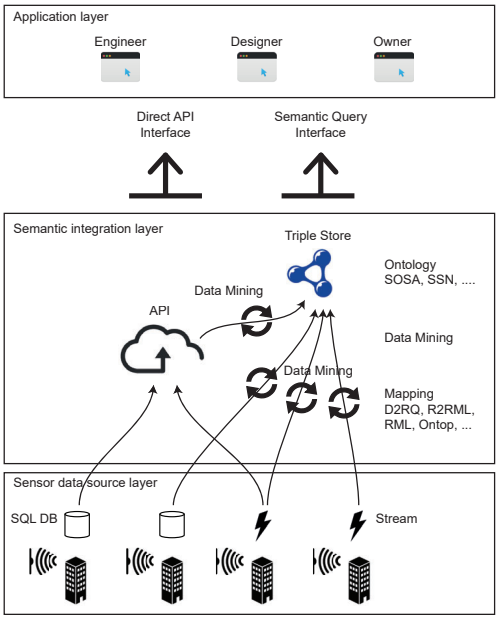

Figure 8: System architecture for the proposed information retrieval platform.

As a result, it is possible to build a decentralised web of semantic information, consisting of various repositories with relevant building data. To maintain that structure and manage the links between the different datasets, we introduce a semantic integration layer with a thin and modular structure (middle in Fig. 8), which captures the semantics of the available data, but at the same time keeps the link to the original data sources in their optimised structures.

5. Implementation

The outlined system for building data representation, pattern mining and decentralised information retrieval has been tested with two use cases. The following sections provide descriptions of the use cases, as well an overview of the system implementation.

5.1. Home2020: Residential building with historical data and no access to real-time data

5.1.1. Use case description

Home 2020 is a detached house completed in 2017 in Denmark (Fig 9) and rated as nearly zero energy building (NZEB) according to the Danish energy labelling standard. The total area of the building is 132 m^2. It consists of a kitchen, a master bedroom, a living room, three other rooms, two bathrooms, a utility room and a walk-in closet. The house is occupied by a young working couple without children.

The heat supply is from district heating, distributed to a floor heating system. The domestic hot water and ventilation with heat recovery (85%) are provided by an air-to-water heat pump integrated in a compact unit. The ventilation system allows individual control of the air supply in the living room and bedrooms and control of the extraction in the kitchen, bathrooms and utility room. The supplied airflow is adjusted in accordance with the CO_2 and relative humidity levels in each room. Automatically controlled natural ventilation grids and skylights aim for enhancing the indoor environmental quality, while simultaneously reducing the energy consumption. The unit is running with a minimum airflow when the house is unoccupied and when a higher air supply is not required by the indoor conditions. The ventilation system is deactivated when the windows and doors are opened.

Figure 9: 3D visualisation of Home2020.

External solar shading devices have been installed in the living room and bedroom ('koekken' and 'Soveværelse' in Fig. 10). Both natural ventilation and shading systems can be controlled automatically based on the temperature, CO_2 level, relative humidity, and occupancy. The control strategy has been implemented towards the end of summer of 2018.

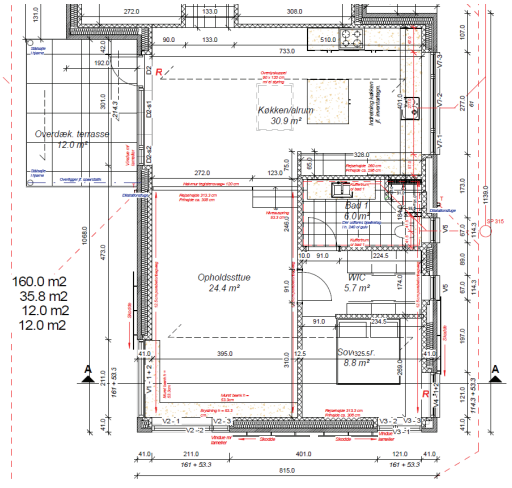

Figure 10: Floor plan of the Home2020 house.

5.1.2. Data monitoring and collection

A BMS is used for data collection with a measurement interval of five minutes. The system monitors several different parameters. Energy consumption is recorded for district heating [MWh], floor heating pump [kWh], ventilation system [kWh], control system [kWh], and kitchen appliances [kWh]. Measurements for the compact unit include outdoor air temperature [°C], return air temperature [°C] , return air relative humidity [percent], hot water temperature [°C], supply air temperature [°C], heat pump temperature [°C], ventilation speed [steps]. Both hot and cold water consumption [m^3] are recorded as well. With regards to indoor environmental quality control, a sensor network distributed over all spaces monitors temperature [°C], CO_2 level [ppm], relative humidity [%], and damper opening [min/ max]. The collected dataset is from the period 01.12.2017 to 31.10.2018. All data is provided in CSV files, with each CSV file containing the sensor data for one day (335 log files in total). An extract of the available data is given in Table 1. Rows contain the measurement points in time and the columns contain the different sensor observations (76 distinct observed variables).

Table 1: Extract of available sensor observations for Home2020: Temperature, CO_2, Relative Humidity.

Time	Temp. (° C)	CO_2 (ppm)	RH (%)
01/01/2018 00.00.47	23.0	742.0	42.0
01/01/2018 00.05.45	23.0	746.0	42.0
01/01/2018 00.10.45	23.0	732.0	42.0
01/01/2018 00.15.45	23.0	738.0	42.0
01/01/2018 00.20.45	23.0	732.0	41.0

For this study, we considered only the sensor data related to indoor environmental quality control, i.e. temperature [°C], CO_2 level [ppm], and relative humidity [%]. These observed variables are available for the entire period and for all rooms. In this article, we present our findings from the kitchen, bedroom and living room in three different months: January, April and August. The choices reflect highest variability in room function and occupant behaviour, as well as external conditions (e.g. seasonal changes in the weather).

5.1.3. Transforming time series data into SAX representations

The fist step in the pattern mining process consists of loading all data into long collections of *Measurements*, with each *Measurement* containing a *Datetime* stamp and a set of *Property* values. Data cleansing and preparation for mining are of utmost importance to the results. That includes performing multiple imputation for removal of missing values and potentially discarding parts of the dataset. In this case, the sensor data values for the period between 26-31 October were discarded, as they did not contain correct and complete data. After the necessary preparatory steps, 94434 measurements in total are parsed and loaded.

Second, Symbolic Aggregate Approximation (SAX) is applied on the loaded measurement values. Namely, pattern mining is not going to be performed directly on the data, but rather on symbolic representations of the data. SAX allows for dimensionality reduction and indexing with a lower bounding distance measure [95]. As defined by Lin et al. [95], to reduce the time series from n dimensions to w dimensions, the data is divided into w segments, and each segment is replaced by the average of its data points (Piecewise Approximate Aggregation (PAA)). The value of each segment is then replaced by a symbol, as the number of symbols and segments is decided by the analyst.

In this research effort, all time series were processed at once and SAX representations were generated with the number of segments equal to 7,869, and the number of symbols equal to seven. The number of segments was set to 7,869 in order to obtain hourly SAX representations. The computation of SAX representations is implemented using the SPMF open-source data mining mining library[13]. The available *Measurement* data are handed to the SPMF tool chain, more particularly to the SAX algorithm[14]. Data is provided per observed variable and for the complete time span. The number of symbols, which was set to seven, takes into account *min* and *max* values over that entire span. Deciding on the number of SAX symbols is a task for the data analyst, and therefore it has the potential to affect the final results. In this case, setting the maximum number of SAX symbols to seven is based on an analysis of the variance in *min* and *max* measured values. For instance, considering the temperature values for the bedroom given in Listing 2, one can easily observe that the difference between the *min* and *max* value for all measurements is approximately four degrees. Setting the number of SAX symbols determines the granularity of the results and has to be justified for all observed variables. And while one may argue that fewer symbols may have provided enough insight into that particular room and observed variable, that may not apply to other rooms and other observed variables. Therefore, based on screening of the general behaviour of all observed variables in all rooms, seven was selected as the number of SAX symbols that would satisfy the interval division criteria for all spaces and observed variables.

```
1 [-Infinity,22.86950723073572]
2 [22.86950723073572,23.704365409749624]
3 [23.704365409749624,24.355554789380466]
4 [24.355554789380466,24.956652678270476]
5 [24.956652678270476,25.60784205790132]
6 [25.60784205790132,26.442700236915222]
7 [26.442700236915222,Infinity])
```

Listing 2: SAX symbols representing the measurement values for the temperature in the bedroom.

Naturally, the symbols for the different rooms and observed variables, despite their equal numerical expression, are different, because they represent different observations and therefore correspond to different interval steps. All SAX symbols are stored in memory, after which an output in TXT files is generated. The complete sequence of data points is replaced by a symbolic representation such as *3222223222222223333*... for the temperature values in the bedroom (first 20 values). Finally, from this data, a matrix indicating the co-occurrence of

[13] http://www.philippe-fournier-viger.com/spmf/

[14] http://www.philippe-fournier-viger.com/spmf/SAXTimeSeries.php

the SAX symbols on a per month basis is computed. A small extract of this matrix is shown in Table 2.

Table 2: Matrix of SAX representations for time series data.

Temperature_Bedroom	3	2	2	2	2	2	2
CO2_Bedroom	4	3	3	1	3	7	7
RH_Bedroom	3	3	4	7	7	5	5
Temperature_LivingRoom	2	2	2	2	2	2	2
CO2_LivingRoom	6	5	5	2	3	6	7
RH_LivingRoom	3	3	4	7	7	7	6
Temperature_Kitchen	2	3	3	3	3	3	3
CO2_Kitchen	6	5	4	1	2	7	7
RH_Kitchen	3	4	4	7	7	5	5

5.1.4. Pattern mining

In the following step, the data is processed to retrieve the frequent repetitive patterns (motifs). This is done by identifying the Longest Repeated Substrings (LRS) in the strings of symbols. In this project, all SAX symbols obtained through the previous step are provided as input, month per month and observed variable per observed variable, to obtain the LRS. The LRS are identified using a custom implementation of the Suffix Tree algorithm. This algorithm identifies and writes all repeated substrings in 27 'lrs.txt' files, as displayed in Listing 3. These 27 files represent the three selected rooms (kitchen, bedroom and living room), for each observed variable (CO_2, temperature and relative humidity), for the selected three months (January, April and August).

```
345555 - 3 - 13;103;130;
444333 - 5 - 78;167;196;504;559;
4445555 - 4 - 29;178;244;642;
44544 - 3 - 124;222;241;
455555556 - 3 - 14;246;598;
455556 - 4 - 31;131;180;644;
54433 - 4 - 62;224;363;432;
55544 - 6 - 107;160;191;217;361;636;
555666 - 10 - 133;147;182;251;309;382;603;621;646;690;
6555554 - 3 - 157;188;723;
66655 - 8 - 141;155;186;301;428;629;681;706;
6667666 - 3 - 137;297;386;
```

Listing 3: An example of LRS found in the SAX sequences. This includes (1) the pattern of SAX symbols, (2) the number of times each appears per month, observed value, and room; and (3) the index in the sequence where the pattern starts.

The output contains some overlapping and redundant patterns, e.g. patterns contained in each other (33334, 333334, etc.). Considering that this effort aims to identify only disjoint, non-redundant and non-overlapping patterns, a manual data cleansing step is included in this point of the pattern mining process to remove redundant data. That is done according to defined criteria, which consider the "interestingness" of the patterns, including length, frequency and evolutionary character. This manual step results in a cleaned set of patterns, stored in 27 distinct files (per room, observed variable, and month).

5.1.5. Finding co-occurrences

The resulting patterns are used to compute the co-occurrence matrices that show which patterns co-occur at any moment in time. In this case, co-occurrence matrices are built per room and month, thereby considering the same three rooms and months. Each co-occurrence matrix thus tracks co-occurring patterns between the observed variables temperature, relative humidity, and CO_2.

The identification of co-occurrences is done in two ways. In a first method, each of the 9 co-occurrence matrices is exported to a list of comma-separated values, containing the ID of the pattern every time a pattern occurs. Figure 11 shows the first part of the output for the patterns in the bedroom in month 8 (August). Figure 12 and Fig. 13 show the bedroom in months 1 (January) and 4 (April) respectively. These data are also visualised in heat maps to ease the detection of co-occurrence of patterns and better understand the pattern distribution throughout the different sequences and therefore the general operational behaviour of the spaces in question. One can see how the different colour segments in the heat map represent the different patterns and match the pattern IDs in the corresponding table in Fig. 11. Listing 4 gives an example overview of the corresponding patterns and their pattern IDs in Fig. 11.

```
332: 77766666
331: 66677
214: 43334
210: 333444
219: 46666666
283: 44544
284: 44566
289: 56655
```

Listing 4: Overview of all the patterns included in Table 11.

As seen in Listing 4, pattern 332 (77766666) is composed by SAX symbols 7 and 6, which happen consecutively in a pattern that appears multiple times within the same SAX sequence. As one can see in the pattern tables, some of these patterns overlap within the same SAX sequence. For example, pattern 332 (77766666) and 331 (66677) overlap from time stamp 24 to 26. That may be helpful in filtering out disjoint patterns, which was previously stated as a criterion. From this visualisation, one can also (manually) find that pattern 332 and 289 appear simultaneously and thus constitute a co-occurrence. Many other motifs can be found in the co-occurrence matrices and heat map visualisations. Yet, as each co-occurrence matrix contains 730 columns (time stamps), it would be inefficient to do such exploration manually. Therefore, a second approach was used as well, which automated the above procedure as much as possible. In the second approach, co-occurrence matrices are composed and calculated in memory, thereby extending the software documented earlier. All patterns are parsed for the entire period, thereby using the pattern identifiers that were already composed. Complete matrices are built in memory, thereby taking into account that multiple patterns can co-occur within the same SAX sequence.

After composing all matrices in memory, each of them is 'stepped through', one datetime value at a time. Each time two patterns co-occur, a co-occurrence object is created, which tracks two co-occurring patterns and the moment when they co-occur. All co-occurrences are listed in memory. Using this list, the co-occurrence matrix is again 'stepped through', and for

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
TEMP (SAX)	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	6	6	6	6	6	7	7	7	7
CO2 (SAX)	4	3	3	3	4	4	4	4	4	4	4	4	4	6	6	6	6	6	6	6	6	5	3	3	3	3	4	4	4	4
HUM (SAX)	5	4	3	3	2	2	2	2	1	2	3	3	3	3	3	4	4	5	4	4	5	6	6	5	5	4	3	4	5	5
TEMP (Pattern)																			332	332	332	332	332	332	332	332	331	331		
CO2 (Pattern)	214	214	214	214	214	210	210						219	219	219	219	219	219	219	219				210	210	210	210	210	210	
HUM (Pattern)																283	283	283	284	284	289	289	289	289	289					

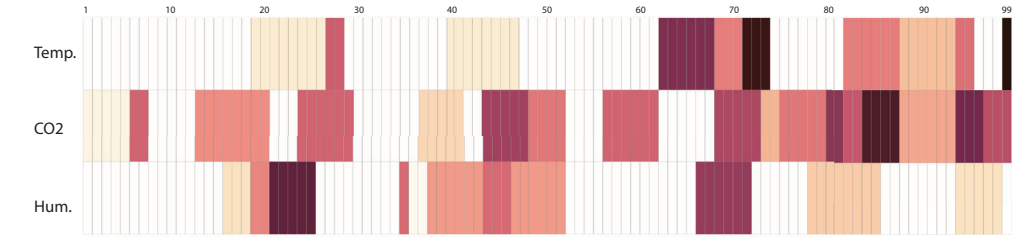

Figure 11: All patterns in the first 100 *Datetime* points visualised in a heat map for the bedroom in August.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
TEMP (SAX)	3	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	2	2	2	2	2	2
CO2 (SAX)	4	3	3	1	3	7	7	7	6	5	4	4	4	4	3	2	3	3	3	3	2	3	3	3	3	3	2	2	3	6
HUM (SAX)	3	3	4	7	7	5	5	6	5	5	5	5	5	4	5	7	6	6	6	6	7	6	6	6	6	7	7	7	7	5
TEMP (Pattern)	307	307	307	307	307	307	307	307	306	306	306	306	306	306	306	306	306	306	298	304	304	304	304	304	304	304	304	304	304	
CO2 (Pattern)						174	175	175	175	175	175	175	175	175	152	152	153	153	153	153	153	152	153	153	153	153	153			
HUM (Pattern)								236	236	236	236	236											237	237	238	241	241	241	241	241

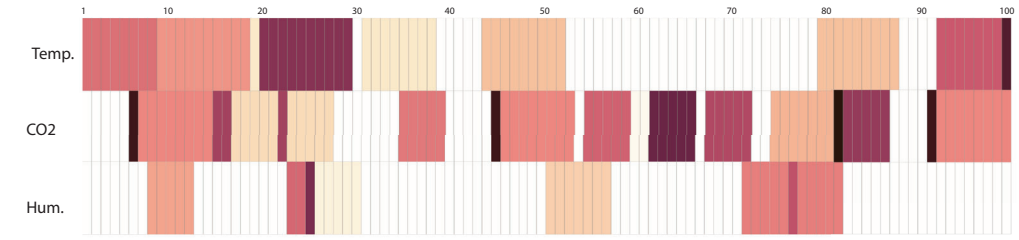

Figure 12: All patterns in the first 100 *Datetime* points visualised in a heat map for the bedroom in January.

each pattern ID encountered, a bag of co-occurring patterns is composed, resulting in ordered lists of co-occurring patterns. At this point, we evaluate to what extent two patterns co-occur through overlap of patterns between the sequences. If overlap of pattern A with pattern B is higher than a certain threshold value (set to 50% for a strong rule of co-occurrence), pattern A overlaps with pattern B, and this co-occurrence is tracked. If 50% of one pattern overlaps with another one, it is not automatically the case that 50% of the other pattern also overlaps with the first one. Thus, each co-occurrence has a source and target pattern, implying direction of the overlap. For the earlier considered time frame (Fig. 11), the co-occurring patterns listed in Listing 5 can be found.

```
332 (77766666): 284, 289
331 (66677): 210
214 (43334): -
210 (333444): -
219 (46666666): 283
210 (333444): 332, 331
283 (44544): 219
284 (44566): 332
289 (56655): 332
```

Listing 5: Co-occurring patterns in the considered time frame.

Initially, the above described algorithm only generates co-occurrences in a pair-wise manner (only two co-occurring items at a time). Although that is highly useful, we need to take into account co-occurrences that include more than two patterns. Based on the computed bags of co-occurrences, multiple co-occurrences are computed as well. This is done in a similar method, the main difference being that triplets are constructed. For each co-occurrence, density of the co-occurrence is computed and traced. If a co-occurrence consists of two patterns, density of co-occurrence can be either 1 or 2 (overlap of 50% in one or two directions, respectively); if a co-occurrence consists of three patterns, density of co-occurrence can be anything from 3 to 6 (overlap of 50% between all three of the included patterns). This continues as the co-occurrences consist of more than 3 co-occurring patterns. No triplets can be found in the bedroom in August (Fig. 11), but an example triplet can be found in the bedroom in January (Fig. 12), namely 304, 153, 237, with a density of 6.

5.1.6. Association rule mining (ARM)

From the bags of co-occurrences in memory, a number of output files are generated, i.e. one file per month for each room

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
TEMP (SAX)	5	5	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
CO2 (SAX)	4	4	4	3	3	3	3	3	5	5	5	5	5	7	7	7	6	6	6	6	5	4	4	4	4	4	3	4	6	7
HUM (SAX)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TEMP (Pattern)	318	318	318	318	318	318	318																							
CO2 (Pattern)	193	193	193	193	193	193									200	200	200	200	200	200	198	197	197	196	196					
HUM (Pattern)																														

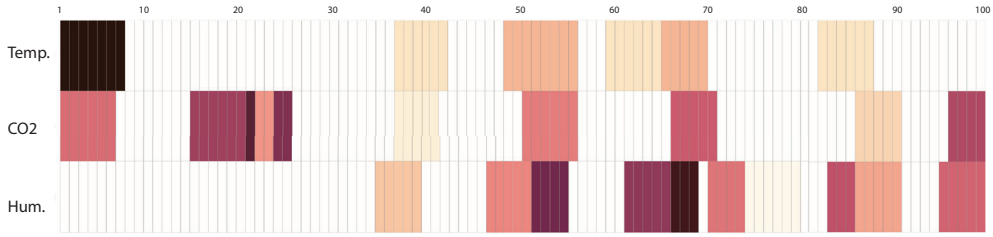

Figure 13: All patterns in the first 100 *Datetime* points visualised in a heat map for the bedroom in April.

for each observed variable (nine files in total). Each file contains one co-occurrence per line, including either two or three patterns per line. These files serve as input for the Association Rule Mining (ARM) step that is executed next, in which each line in each file is considered a 'Transaction' and the totality of all files constitutes the 'Transaction database' required for the rule mining. In mining for association rules, we used the SPMF open source library again[15], which includes an implementation of the FP-growth algorithm. The output of the algorithm consists of the targeted association rules, including the measures of "interestingness" support and confidence. Listing 6 shows a part of the association rules that have been obtained from indoor environmental quality data in the living room in August. Several hundred association rules are discovered in total, the majority of which in the data from the living room and the kitchen.

```
452 ==> 489 #SUP: 1 #CONF: 1.0
453 ==> 485 #SUP: 3 #CONF: 0.6
454 ==> 481 #SUP: 1 #CONF: 0.5
456 ==> 484 #SUP: 2 #CONF: 0.6666666666666666
457 ==> 488 #SUP: 1 #CONF: 1.0
459 ==> 481 #SUP: 1 #CONF: 0.5
459 ==> 488 #SUP: 1 #CONF: 0.5
482 ==> 460 #SUP: 1 #CONF: 0.5
460 ==> 482 #SUP: 1 #CONF: 0.5
460 ==> 485 #SUP: 1 #CONF: 0.5
457 488 ==> 378 #SUP: 1 #CONF: 1.0
378 488 ==> 457 #SUP: 1 #CONF: 0.5
378 457 ==> 488 #SUP: 1 #CONF: 1.0
457 ==> 378 488 #SUP: 1 #CONF: 1.0
459 488 ==> 378 #SUP: 1 #CONF: 1.0
378 488 ==> 459 #SUP: 1 #CONF: 0.5
378 459 ==> 488 #SUP: 1 #CONF: 1.0
459 ==> 378 488 #SUP: 1 #CONF: 0.5
```

Listing 6: Some of the association rules obtained for the living room in August.

As not all rules will be interesting and provide novel insights, further selection needs to be made. A starting point would be the selection of strong rules only, i.e. use only association rules with confidence of 1.0. Further considerations in terms of combined effect of support and confidence measures, as well as prioritising multiple co-occurrences are in the hands of the domain expert/analyst and depend on the purpose of the knowledge discovery. It is important to note that this research effort focuses on knowledge discovery, representation and retrieval from a computational perspective, but the actual interpretation of the discovered patterns and rules in terms of building performance is beyond the scope of this article.

[15] http://www.philippe-fournier-viger.com/spmf/AssociationRules.php

5.1.7. Semantic data modelling

According to the suggested system architecture (Fig. 8) and the introductory sections, the relevant building data and discovered knowledge need to be made accessible to the end users to enable holistic design decision support. Ideally, each type of data is served in the best possible format and datasets are linked across domains. Considering the emergence of semantic web and data modelling techniques worldwide, these technologies are the ideal candidates for such representation. The Home2020 use case building was therefore modelled accordingly, thereby adopting suggested ontologies from the LBD community group. For reference, Listing 7 shows all namespaces and URIs (Unique Resource Identifiers) used in this effort.

```
@prefix seas: <https://w3id.org/seas/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix bot: <https://w3id.org/bot#> .
@prefix geo-ext: <http://eapetrova.com/voc/geoextension#> .
@prefix bmeta: <http://eapetrova.com/voc/buildingmetadata#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix ssn: <http://www.w3.org/ns/ssn/> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix om: <http://www.ontology-of-units-of-measure.org/resource/
  om-2/> .
@prefix ptn: <http://eapetrova.com/pattern/> .
@prefix list: <https://w3id.org/list#> .
@prefix inst: <https://home2020.dk/instances#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
```

Listing 7: All namespaces used in the RDF graph.

As a start, the building itself has been modelled as an RDF graph according to the BOT ontology (Listing 8). This graph contains the description of building, building storeys and spaces. Also latitute, longitude, and altitude of the building

are included using geospatial ontologies, as well as an OpenStreetMap (OSM) location[16]. The `ssn:hasProperty` predicate links each of the spaces to the sensor observations that are measured inside. Furthermore, the `bot:containsElement` containment relation relates the space to its contained sensor node.

```
inst:Home2020BuildingSite
  rdf:type owl:NamedIndividual, bot:Site ;
  rdfs:label "Site of the building"@en ;
  bot:hasBuilding inst:BuildingHome2020 .

inst:GroundFloor
  rdf:type owl:NamedIndividual, bot:Storey ;
  rdfs:label "Ground floor of the building"@en .

inst:BuildingHome2020
  rdf:type owl:NamedIndividual, bot:Building;
  rdfs:label "Passive house"@en;
  bot:hasStorey inst:GroundFloor ;
  bot:hasSpace inst:Kitchen , inst:LivingRoom , inst:Bedroom ;
  geo:lat "56.0914290" ;
  geo:long "9.7958060" ;
  geo:alt "16" ;
  geo-ext:inOSMLocation <https://www.openstreetmap.org/node
    /3721416569> .

inst:Kitchen
  rdf:type bot:Space, sosa:FeatureOfInterest ;
  bot:containsElement inst:sensorNode_1 ;
  rdfs:label "Kitchen"^^xsd:string ;
  ssn:hasProperty inst:Kitchen-CO2, inst:Kitchen-Temperature, inst:
    Kitchen-Humidity .
```

Listing 8: RDF graph for the Home2020 building.

The sensor nodes are furthermore linked to their corresponding sensor observations. As indicated in Section 2.2.3, these sensor observations can be represented using the SSN and SOSA ontologies. This results in data as shown in Listing 9. Sensor nodes are linked to the individual sensors (`sosa:hosts`); for each sensor, an indication is given of what it observes (`ssn:observes`); and the `sosa:madeBySensor` predicate links each of the observations (e.g. `inst:Kitchen-CO2-Sensor-obs1`) to its corresponding sensor. For each observation, numerical measures, units and datetime of measurement are included using SOSA and OM (Units of Measure) ontologies.

```
inst:sensorNode_Kitchen
  rdf:type sosa:Platform ;
  sosa:hosts inst:Kitchen-CO2-Sensor, inst:Kitchen-Temperature-
    Sensor, inst:Kitchen-Humidity-Sensor ;
  ptn:hasAssociationRule inst:associationRule_1, inst:
    associationRule_2 .

inst:Kitchen-CO2
  rdf:type sosa:ObservableProperty .

inst:Kitchen-CO2-Sensor
  rdf:type sosa:Sensor ;
  ssn:observes inst:Kitchen-CO2 .

inst:Kitchen-CO2-Sensor-obs1
  rdf:type sosa:Observation ;
  sosa:hasFeatureOfInterest inst:Kitchen ;
  sosa:hasResult [ a om:Measure ;
  om:hasNumericalValue "809.0"^^xsd:double ;
  om:hasUnit om:partsPerMillion ] ;
  sosa:madeBySensor inst:Kitchen-CO2-Sensor ;
  sosa:observedProperty inst:Kitchen-CO2 ;
  sosa:resultTime "01/12-2017 00:00:47"^^xsd:dateTime .

inst:associationRule_1
  rdf:type ptn:AssocationRule ;
  ptn:LHS (inst:Motif_45) ;
  ptn:RHS (inst:Motif_137) ;
  ptn:confidence "0.5"^^xsd:double ;
  ptn:absoluteSupport "1"^^xsd:double ;
  ptn:relativeSupport "0.5"^^xsd:double .

inst:motif_45
  rdf:type ptn:Motif ;
  ptn:SAXsequence "11122"^^xsd:string ;
  ptn:space inst:Kitchen ;
  ptn:month "8"^^xsd:string ;
  ptn:SAXsequenceFull (inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7
    b177e237bc0 inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7b177e237bc0
     inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7b177e237bc0 inst:
    SAXSymbol_41fadfdb-6560-4e96-9a7f-bc405f453452 inst:
    SAXSymbol_41fadfdb-6560-4e96-9a7f-bc405f453452 );
  ptn:observedVariable "CO2"^^xsd:string .

inst:SAXSymbol_36ef82d8-57c9-4e0a-a0bc-c1c66404b02b
  rdf:type ptn:SAXSymbol ;
  ptn:symbol "5"^^xsd:int ;
  ptn:lowerBound "645.651281059915"^^xsd:double ;
  ptn:upperBound "700.959674546294"^^xsd:double .
```

Listing 9: RDF graph for the Home2020 building.

Finally, the graph also contains all association rules and motifs found in the data (see Section 5.1.2 to 5.1.6). These are stored in the graph using a built-for-purpose PATTERN ontology (`ptn:`). This ontology allows to represent the discovered association rules, including their `ptn:confidence`, `ptn:absoluteSupport`, and `ptn:relativeSupport`. These association rules (`inst:associationRule_1`) are linked to individual sensor nodes using `ptn:hasAssociationRule` predicates. Furthermore, the association rules link to ordered lists of motifs (patterns) on the left-hand side (`ptn:LHS`) and right-hand side (`ptn:LHS`) of the rule. These motifs are documented in the graph as well (e.g. `inst:motif_45`), including its corresponding SAX symbols (e.g. `11122`), month and space in which the pattern was found, and full representation of each of the linked SAX symbols (lower bound, upper bound, symbol).

5.2. Gigantium: Cultural and sports centre with historical data and access to real-time data

A second use case has been documented for the Gigantium building, where besides access to historical data, there is also access to continuous incoming streams of real-time monitored data. This use case follows largely the same approach as the Home2020 case. Therefore, this section will mainly highlight the differences compared to the previous case and the obtained results.

5.2.1. Use case description

Gigantium is a large cultural and sports center in Aalborg, Denmark, which opened in 1999. At that time, it housed only an indoor football and handball hall, a sports hall and meeting rooms. Two ice skating rinks were added in 2007 together with swimming pool and wellness areas in 2011. Currently, Gigantium consists of an ice skating arena, ice rink for training purposes, sports halls, a concert and exhibition halls, swimming pool and wellness area, athletics hall, conference rooms, a cafe, and a visitors lobby. The total area of the center is 34000 m^2.

[16] https://www.openstreetmap.org/

The maximum capacity of the ice skating arena is 5000 spectators and that of the main hall during concerts is 8500.

Operational building data is being collected through a sensor network consisting of 39 nodes, divided between the spaces. The placement of the sensor nodes is indicated in the yellow numbered rectangles in Fig. 14. The sensors monitor Temperature (°C), Relative Humidity (%), Air Pressure (hPa), Indoor Air Quality [Total Volatile Organic Compounds ((TVOC), ppb) and CO_2 (ppm)], illuminance (lux), motion and noise levels. The data collection serves multiple purposes, including monitoring indoor climate and thermal comfort for the visitors and providing information on space use for the facility management staff. When it comes to behavioural insights into the building operation, the diversity of facilities and related activities will definitely have an effect, as no uniform building or occupant profile will be possible. For example, the temperature and relative humidity in the meeting rooms, ice hockey arenas, fitness and wellness areas will differ significantly. Thus, this use case can be used to test the proposed knowledge discovery approach in diverse environments and provide multifaceted behavioural and query results.

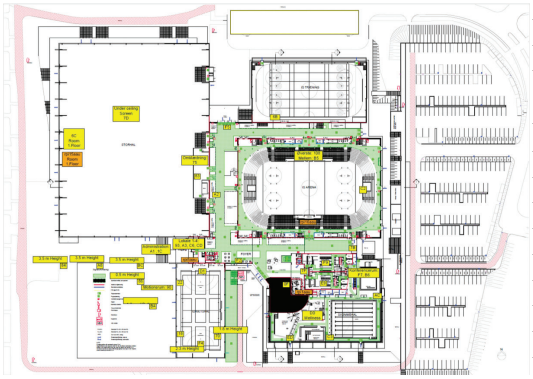

Figure 14: Floor plan of the Gigantium building with an indication of sensor nodes in yellow rectangles.

In both cases (Home2020 and Gigantium), the discovered motifs and association rules can be used to inform design decisions related to spatial design, thermal comfort, indoor climate and HVAC system design. For example, the case of Gigantium presented significant issues related to overheating in the conference room, which led to a decision to renovate the mechanical ventilation system. The discovered insights would be of high value to the decision-making related to the new system design and can help prevent uninformed decisions or reuse of inaccurate design parameters that previously led to this underperformance.

5.2.2. Knowledge discovery

The analysed dataset is collected between March and May 2018. All repeated pattern instances in the symbolic representation of the time series were identified, following the same approach and criteria as in the previous use case. Figure 15 shows a visual representation of the labelled discovered 14 motifs (M1, M2, ..., M14) in the sequence of the six variables for the visitors' cafe. Clearly, the smaller size of the dataset and the profile of the space are reflected in the much smaller amount of discovered patterns.

To enable association rule mining, the discovered motifs are further used to construct a co-occurrence matrix. Using the co-occurrence matrix, we obtained 10 sets of co-occurring items for the considered period and space. After performing the association rule mining, we discover 13 strong association rules (i.e. confidence equal to 1). Nine association rules are related to the co-occurrence of M7, M9 and M14. Other association rules are M1 => M10, M3 => M10, M12 => M10, M13 => M8, where M8 => M13 is identified as a bidirectional association rule. In this case, the rule indicates an association between observation patterns related to air pressure and CO_2. As previously mentioned, the meaning of the rules needs to be interpreted relatively to the knowledge discovery and decision support purposes.

Once again, to be able to reuse the discovered knowledge, it also has to be represented accordingly and connected to the semantic graph. This is done in a way similar to the Home2020 case, by modelling the rules and linking this graph to the representation of the space hosting sensor node 00000014, to create the motif-enriched graph.

5.2.3. Semantic data modelling

Similarly to the previous use case, the spaces are represented using the BOT ontology as `bot:Space` instances and then linked to the corresponding hosted sensor nodes represented with the SOSA ontology. Each sensor node hosts sensors, tracking the six observed variables. Besides using SOSA to model the sensor nodes and their metadata, in this case we also use a separate ontology with prefix `bmeta:`, to model the measurements associated to each sensor. Most importantly, in contrast to the Home 2020 case, the data values are not directly stored in the semantic graph. Instead, a custom `bmeta:values` datatype property points to a web address that returns the data values as requested using the HTTP protocol (see Listing 10). As documented previously in Petrova et al. [10], it is possible to add attributes to the HTTP requests, thereby setting query parameters such as time frame and refresh rate (e.g. from=now-30d&to=now&refresh=30s). The result includes the pointer to the data stream for a `sosa:Result` of a `sosa:Observation`. However, external access to the sensor data streams is obviously restricted.

```
inst:room_1
    rdf:type bot:Space ;
    rdfs:label "Main hall" ;
    bot:hasSpace inst:room_2 ;
    bot:containsElement inst:sensorNode_00000097, inst:
      sensorNode_000000B0, inst:sensorNode_00000077 ;
    geom:hasGeometry "2000, 3000, 4000, 6000"^^wkt:linestring.

inst:sensorNode_00000097
    rdf:type sosa:Platform, bot:Element ;
    rdfs:label "00000097" ;
    bmeta:observation "Space use" ;
    sosa:hosts inst:sensor_00000097_1 ;
    bmeta:placement "Placed in the middle of the hall, 8m above the
      floor."
```

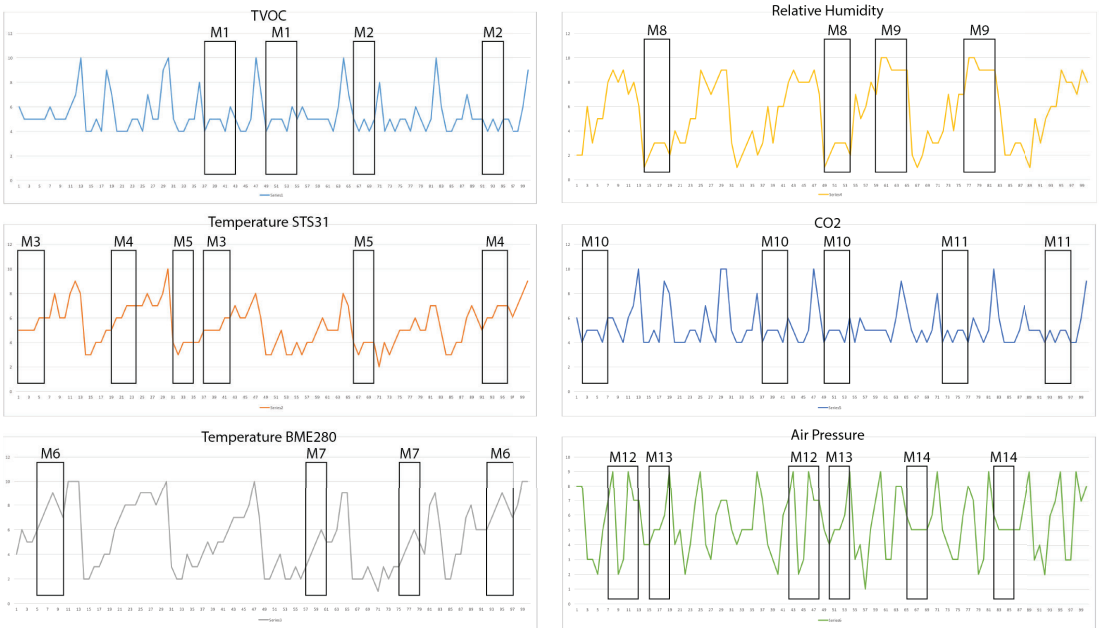

Figure 15: Overview of the co-occurring motifs that have been discovered in the indoor environmental quality data in Gigantium.

```
inst:result_1
    rdf:type sosa:Result ;
    rdfs:label "Result of observation of Relative Humidity" ;
    bmeta:values "https://gigantium.dk/Gigantium2018instances?orgId
      =1&datastream=true" .
```

Listing 10: RDF graph for the Gigantium building.

6. Information retrieval results

In a final test of the proposed approach towards the combination of semantic data modelling and KDD for design decision support, this section looks into the retrieval of the discovered knowledge in the design environment. We demonstrate how an active design case can be connected to a repository of design data enriched with patterns and rules obtained using the KDD process.

6.1. Building data repository

To achieve optimal results, the information retrieval should start from a rich knowledge base hosting heterogeneous data from diverse buildings. Such knowledge bases are vital to the performance of the decision support systems, but they are not openly available and take time to build up to a level where they can respond to potentially any query in a fulfilling way. Therefore, to demonstrate the information retrieval, we create the knowledge base using a self-owned collection of 531 building models originally available in the IFC data model. The IFC models are converted to linked data by the use of the IFC-to-LBD converter[17]. The resulting RDF graph and the contained data are easy to query with out-of-the-box languages such as SPARQL.

The resulting semantic graphs in TTL format are compliant with the BOT ontology and further enriched with BuildingElement[18], DistributionElement[19], and PSET ontologies[20]. For the purposes of this study, geometric data is excluded from the conversion to LBD, leaving only the semantic backbone and product data for each building model. Geometric data may, of course, also be converted to linked data and added to the graph, but this would be less useful for semantic information retrieval in this case, as 3D geometry and BIM models are not available for the existing use case buildings (Home2020 and Gigantium). Furthermore, to be useful for information retrieval, raw geometry should be processed to contain semantically useful concepts (e.g. above, below, next to) which is out of scope for this research effort.

The conversion results in a collection of two Stardog triple stores, with a total of 36 Million triples divided between them. By spreading the data over two stores, we mimic a real-world scenario in which more than one repository is available and needs to be queried using a federated query approach. The data includes 372 `bot:Building` instances, 3,523

[17] https://github.com/jyrkioraskari/IFCtoLBD
[18] https://pi.pauwel.be/voc/buildingelement
[19] https://pi.pauwel.be/voc/distributionelement
[20] http://app.informationdeliveryspecification.org/psets/IFC4/index.html

bot:Zone instances, 2,117 bot:Space instances, and 615,452 bot:Element instances. The bot:Element instances also include a product type (wall, window, etc). The graphs for the use case buildings (Gigantium, Home2020) are added to this repository, including the monitored data, discovered motifs, association rules, etc.

6.2. Matching and query performance

In this article, we limit to an investigation of possibilities for information retrieval, without going in detail on the design environment that is served the retrieved information. Indeed, in a design process, information retrieval needs to be triggered from within the design environment used by the design team. This will likely be a BIM tool, but other tools may be used as well. How the information then gets used, is a research effort in itself and is out of scope here. We do know, however, that information retrieval for decision support from the knowledge base will include SPARQL queries, and we can give an indication of query performance in this section.

In order to obtain reference knowledge from the building data repository, queries will be formed and executed depending on the context of the design team and project. In our case, we envision a recommendation tool setup, in which a design team is working in a BIM environment and would benefit from relevant knowledge present in previous building projects and actively used buildings. In such a case, a key query would be to retrieve buildings or spaces of the same type. In our case, we can use the rdfs:label tags for that purpose. It would be even better, though, if all buildings had the same standardised classification tags used throughout the repository (e.g. Getty AAT tags[21]). Alternative queries to obtain reference buildings and/or spaces are of course also possible. Listing 11 shows an example SPARQL query, which retrieves a list of relevant building and space URIs. This is a federated query, relying on the SERVICE keyword in SPARQL to be able to query both building data repositories at once.

```
SELECT ?b WHERE{
    {
        SELECT ?b WHERE {
            SERVICE <http://localhost:5820/BuildingDataRepo1/query>
                {
                    ?b a bot:Building .
                    ?b bot:hasSpace ?s .
                    ?s rdfs:label "Kitchen"^^xsd:string ;
                }
        }
    }
    UNION
    {
        SELECT ?b WHERE {
            SERVICE <http://localhost:5820/BuildingDataRepo2/query>
                {
                    ?b a bot:Building .
                    ?b bot:hasSpace ?s .
                    ?s rdfs:label "Kitchen"^^xsd:string ;
                }
        }
    }
}
```

Listing 11: SPARQL query for relevant buildings, federated over the available Stardog databases.

[21] http://www.getty.edu/research/tools/vocabularies/aat/

For each of the resulting URIs, relevant knowledge is now available, as displayed in the graph in Fig. 16 for the Home2020 case. This graph shows the BOT topology of the building towards the top of the Figure (Bedroom, Kitchen, Living Room, Site, Building). The 'Passive house' node is identical to one of the building URIs retrieved using the query in Listing 11. As can be seen, this allows to retrieve the sensorNode in the Kitchen (bot:containsElement), for which several association rules are available. Furthermore, the three contained sensors (CO_2, Temperature, Relative Humidity) can be retrieved, including the actual observation measurements (left and bottom of the graph in Fig. 16).

The returned URIs (Listing 11) for spaces and buildings are reference points for obtaining more data. These URIs can be used by the BIM tool to subsequently query for building performance patterns that are available for the retrieved buildings and spaces (Listing 12). As such, it is possible to obtain a graph as displayed in Fig. 17, including ARMs, motifs, and observations.

```
SELECT ?sensor ?ar ?obs WHERE {
  ?s a bot:Space .
  ?s bot:containsElement ?sn .
  ?sn sosa:hosts ?sensor .
  ?sn ptn:hasAssociationRule ?ar .
  ?sensor ssn:observes ?obsp .
    ?obs sosa:hasFeatureOfInterest ?s .
}
```

Listing 12: SPARQL query for patterns and observation values.

The graph in Fig. 17 starts from one association rule, namely associationRule_1, which is linked to one of the sensor nodes (top of Fig.17). This association rule is linked to two motifs (left-hand side and right hand side). The graph allows to retrieve other association rules linked to those motifs. Furthermore, for each motif, the associated SAX representations are available (right in Fig.17), including month and observed variable. By building an appropriate user interface on top of this data, which is out of scope of this research article, appropriate feedback can be retrieved from the building data repository, in support of sustainable BIM-based design.

7. Conclusions

This article investigates the potential of KDD and semantic data modelling for achieving evidence-based sustainable BIM-based building design by establishing a feedback loop from building operation to design. While each of these approaches alone may not be sufficient to close that cycle, we demonstrate that combining them is especially useful for discovery of valuable hidden knowledge in the operation of the existing building stock and documenting it in a reusable, modular and extensible way. This combined approach can help enhance design decision-making and contribute to the improvement of building performance, indoor environmental quality, operation and occupant comfort.

To showcase the above-mentioned potential, we first perform an extensive literature review to identify the available approaches for semantic data modelling and KDD in the AEC

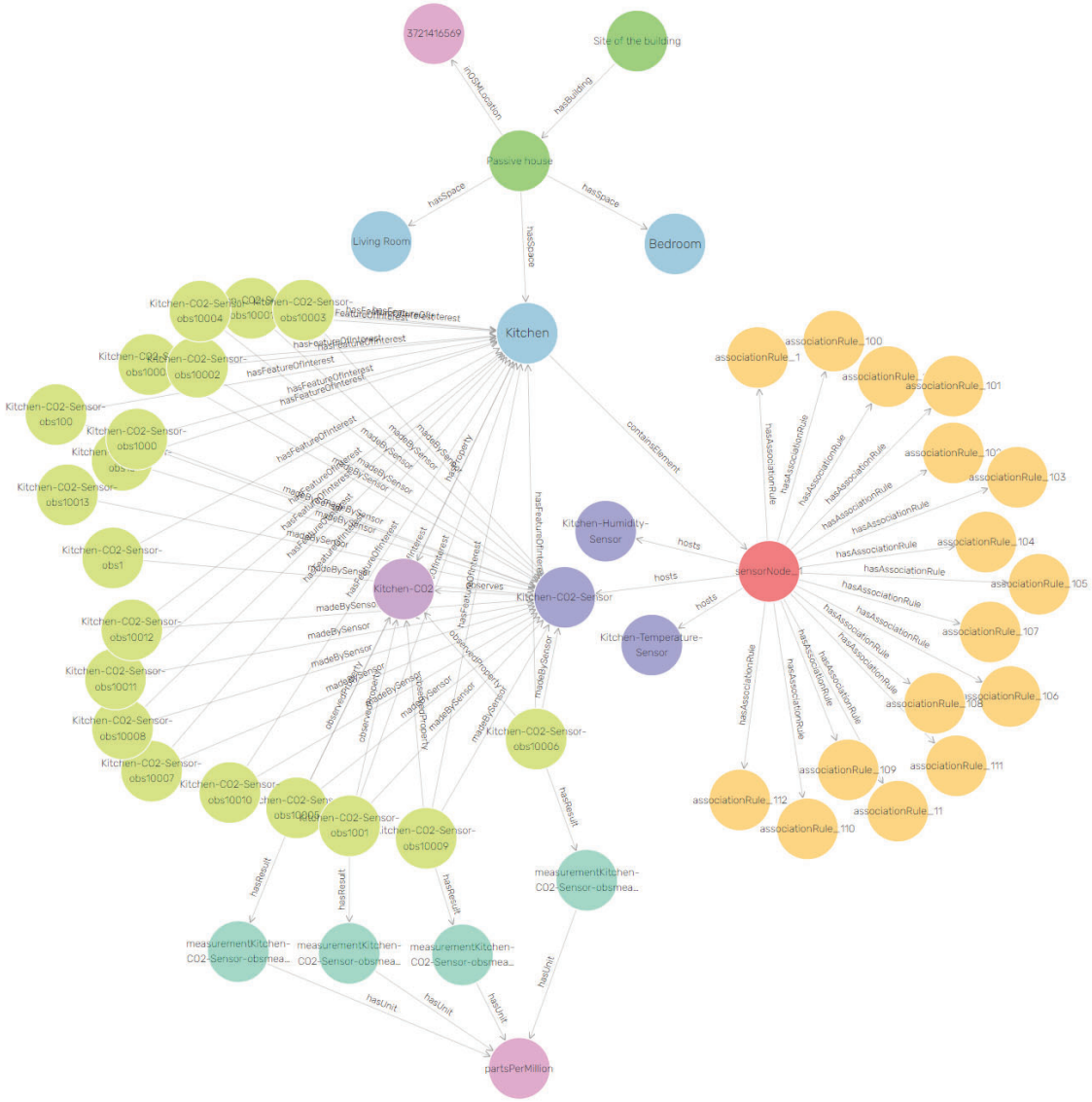

Figure 16: Semantic graph accessible for each of the building URIs obtained using the query in Listing 11.

industry, as well as the contributions that use those to target improvement of decision-making and building performance. Based on that, we outline a system architecture that integrates KDD and semantic technologies for design decision support. The implementation is demonstrated with two use cases, for which motif discovery and association rule mining are performed on the available operational building data, and semantic data modelling is used for representation and retrieval of the discovered knowledge. The resulting knowledge graphs include building data, (links to) the actual sensor data, frequent repetitive patterns and association rules, and thus provide an ideal resource for user-centred design decision support.

Such an approach to building design holds a much bigger potential than performance optimisation. Capturing the material, energy and behaviour metabolisms of buildings and their occupants can provide opportunities for much more sophisticated performance assessments and a higher level understanding of the built environment. Linking the discovered knowledge

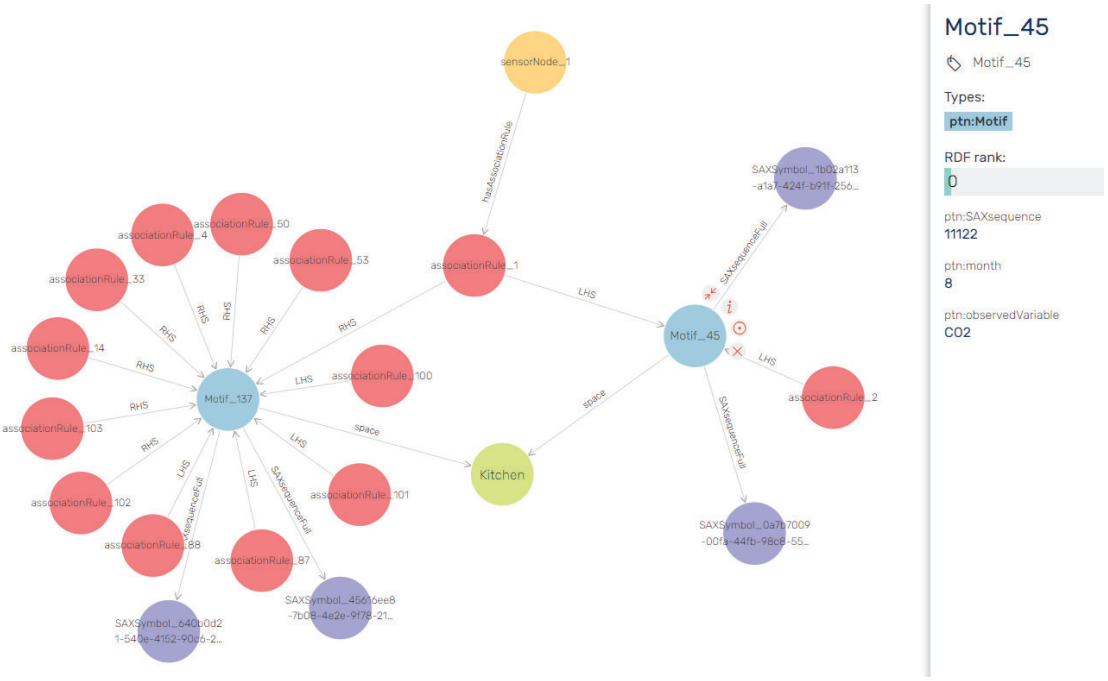

Figure 17: Graph with motifs and ARMs and observation measurements.

to other relevant data sets in a cohesive ecosystem has the potential to revolutionise our understanding of sustainability and help design buildings that adapt to future requirements. The suggested approach allows to redefine the view on the built environment from a technological asset to a higher-level contributor and informant to what the next generation buildings and the related design processes should constitute. For design practice; being able to trace back from discovered performance patterns to the original data will make it possible to always have a dynamic live link between the existing building and its semantic representation.

A number of valuable conclusions can be made from the conducted experiments, which highlight some main challenges and inspiration for future research:

- *Data handling and automation:*
  The Home2020 data is currently available in log files (historical data), which have been extracted from a live system since the data collection from the building continues. This results in information management issues, in the sense that the patterns and rules are separate from the live data. Any future patterns will need to be discovered in a new set of log files. This results in additional manual work, little automation and no real-time performance insight. Even though the graph for Gigantium includes a direct API connection, also in that case, the mining happened on extracts from the live system; hence, certain manual work remains. Future work may look further in automating knowledge discovery directly from the live data (e.g. stream processing and mining of data streams).

- *Interpretation of the discovered knowledge:*
  Even though the semantic graphs contain multiple motifs and association rules, their human interpretation is still required. Hence, there is a need for presenting the discovered knowledge to experts, so that they can semantically annotate the discovered motifs and association rules.

- *Dependency on symbolic approximation choices:*
  In the Home2020 case, each SAX symbol represents one hour of data. As a result, motifs are found for relatively big spans of time. Furthermore, SAX symbols were computed in seven intervals for all observation sequences for the entire period. Also, if temperature values are all between 22 and 23°Celcius, there is little variation, and seven symbols may make less sense. More custom choices could be made (e.g. 14 intervals for January and three intervals for August; 20 minute approximation instead of hourly; etc.). Hence, a careful choice needs to be made for each of the knowledge discovery steps. This is an obstacle towards full automation: high-quality knowledge discovery requires good interplay between manual and automated steps.

- *Support and confidence as measures of significance:*
  At the moment, all discovered association rules are stored

in the graph, including support and confidence values. However, high support and confidence can be a determinative indicator of interestingness and value. Such rules could be filtered at query runtime. This is an important feature, as experts and analysts should be presented with the most important patterns first in their annotation tasks.

- *Swollen graph challenge:*
  It is possible to add sensor data values, rules, motifs, and SAX symbols to the graph, as has been done in the Home2020 case. Yet, this results in a significantly larger graph. For Home2020, the graph size went from 14kB to 414MB because of this step and the demonstration was done for only three rooms. This has an impact on query performance. The swollen graph issue can be prevented by storing such data values in dedicated systems, such as relational databases (sensor data) or binary data formats (images, 3D geometry), and storing a link to those systems in the graph, as has been done for the Gigantium case.

- *Stability of the ontologies:*
  The graphs currently rely heavily on a number of vocabularies (SOSA, BOT, SSN, OM, etc.). This allows to represent several buildings in the same way. It also allows to query across those buildings and their data. If vocabularies change over time, the data also needs to be reformatted accordingly, which may result in data loss. Even if the change over time cannot be prevented, vocabularies should ideally be kept as stable as possible and to some extent standardised across the AEC industry.

8. References

[1] T. Kocaturk, Towards an intelligent digital ecosystem - sustainable data-driven design futures, in: P. Brandon, P. Lombardi, G. Shen (Eds.), Future Challenges in Evaluating and Managing Sustainable Development in the Built Environment, Wiley-Blackwells, Chichester, UK, 2017, pp. 164–178. doi:10.1002/9781119190691.ch10.

[2] A. Borrmann, M. König, C. Koch, J. Beetz, Building Information Modeling: Technology Foundations and Industry Practice, 1 ed., Springer, 2018. doi:10.1007/978-3-319-92862-3.

[3] R. Sacks, C. M. Eastman, G. Lee, P. Teicholz, BIM handbook: a guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators, 3 ed., John Wiley & Sons, Hoboken, NJ, USA, 2018.

[4] P. de Wilde, The gap between predicted and measured energy performance of buildings: A framework for investigation, Automation in Construction 41 (2014) 40 – 49. doi:10.1016/j.autcon.2014.02.009.

[5] E. Corry, P. Pauwels, S. Hu, M. Keane, J. O'Donnell, A performance assessment ontology for the environmental and energy management of buildings, Automation in Construction 57 (2015) 249–259. doi:10.1016/j.autcon.2015.05.002.

[6] S. Hu, E. Corry, E. Curry, W. J. Turner, J. O'Donnell, Building performance optimisation: A hybrid architecture for the integration of contextual information and time-series data, Automation in Construction 70 (2016) 51 – 61. doi:10.1016/j.autcon.2016.05.018.

[7] E. Petrova, P. Pauwels, K. Svidt, R. Jensen, Towards data-driven holistic sustainable design: A decision support framework relying on knowledge discovery in real-time building performance data and disparate project data repositories, Architectural Engineering and Design Management (2018) 1–23. doi:10.1080/17452007.2018.1530092.

[8] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine 17 (1996) 37–54.

[9] E. Petrova, P. Pauwels, K. Svidt, R. Jensen, From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches, in: Proceedings of the 2018 European Conference on Product and Process Modelling (ECPPM), 2018, pp. 391–399.

[10] E. Petrova, P. Pauwels, K. Svidt, R. Jensen, In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data, in: Advances in Informatics and Computing in Civil and Construction Engineering, 2018, pp. 19–27.

[11] G. Piatetsky-Shapiro, Knowledge discovery in real databases: A report on the ijcai-89 workshop, AI Magazine 11 (1991) 68–70.

[12] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, 2011.

[13] J. Han, M. Kamber, J. Pei, Data mining concepts and techniques, 3 ed., Morgan Kaufmann, Waltham, US, 2012.

[14] A. Lausch, A. Schmidt, L. Tischendorf, Data mining and linked open data – new perspectives for data analysis in environmental research, Ecological Modelling 295 (2015) 5–17.

[15] T.-C. Fu, A review on time series data mining, Engineering Applications of Artificial Intelligence 24 (2011) 164–181. doi:10.1016/j.engappai.2010.09.007.

[16] S. Shekhar, P. Zhang, Y. Huang, Spatial data mining, Data Mining and Knowledge Discovery Handbook (2005) 833–851.

[17] Z. J. Yu, F. Haghighat, B. C. Fung, Advances and challenges in building engineering and data mining applications for energy-efficient communities, Sustainable Cities and Society 25 (2016) 33 – 38. doi:10.1016/j.scs.2015.12.001.

[18] C. Fan, L. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review, Energy and Buildings 159 (2017). doi:10.1016/j.enbuild.2017.11.008.

[19] M. Molina-Solana, M. Ros, M. Ruiz, J. Gómez-Romero, M. Martin-Bautista, Data science for building energy management: A review, Renewable and Sustainable Energy Reviews 70 (2017) 598–609. doi:10.1016/j.rser.2016.11.132.

[20] H.-X. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (2012) 3586 – 3592. doi:10.1016/j.rser.2012.02.049.

[21] Z. Wang, R. Srinivasan, A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models, Renewable and Sustainable Energy Reviews (2016). doi:10.1016/j.rser.2016.10.079.

[22] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205. doi:10.1016/j.rser.2017.04.095.

[23] T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review, Energy and Buildings 165 (2018) 301–320. doi:10.1016/j.enbuild.2018.01.017.

[24] K. Li, X. Xie, W. Xue, X. Dai, X. Chen, X. Yang, A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction, Energy and Buildings 174 (2018) 323 – 334. doi:10.1016/j.enbuild.2018.06.017.

[25] W. Kim, S. Katipamula, A review of fault detection and diagnostics methods for building systems, Science and Technology for the Built Environment 24 (2018) 3–21. doi:10.1080/23744731.2017.1318008.

[26] A. Capozzoli, M. Piscitelli, S. Brandi, D. Grassi, G. Chicco, Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings, Energy 157 (2018) 336–352. doi:10.1016/j.energy.2018.05.127.

[27] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big bas data for building energy management, Energy and Buildings 109 (2015) 75–89.

[28] C. Fan, L. Xiao, Y. Zhao, J. Wang, Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data, Applied Energy 211 (2018) 1123–1135. doi:10.1016/j.apenergy.2017.12.005.

[29] M. Pena, F. Biscarri, J. Guerrero, I. Monedero, C. León, Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach., Expert Systems with Applications 56 (2016). doi:10.1016/j.eswa.2016.03.002.

[30] S. Fong, J. Li, W. Song, Y. Tian, N. Dey, Predicting unusual energy con-

sumption events from smart home sensor network by data stream mining with misclassified recall, Journal of Ambient Intelligence and Humanized Computing 9 (2018). doi:10.1007/s12652-018-0685-7.

[31] N. Zhu, A. Anagnostopoulos, I. Chatzigiannakis, On mining IoT data for evaluating the operation of public educational buildings, in: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2018, pp. 278–283.

[32] C. Fan, Y. Sun, K. Shan, F. Xiao, J. Wang, Discovering gradual patterns in building operations for improving building energy efficiency, Applied Energy 224 (2018) 116 – 123. doi:10.1016/j.apenergy.2018.04.118.

[33] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, J. Wang, A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning, Applied Energy 235 (2019) 1551 – 1560. doi:10.1016/j.apenergy.2018.11.081.

[34] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renewable and Sustainable Energy Reviews 81 (2018) 1365 – 1377. doi:10.1016/j.rser.2017.05.124.

[35] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Automation in Construction 49 (2015) 1 – 17. doi:10.1016/j.autcon.2014.09.004.

[36] M. Ashouri, F. Haghighat, B. C. Fung, A. Lazrak, H. Yoshino, Development of building energy saving advisory: A data mining approach, Energy and Buildings 172 (2018) 139 – 151. doi:10.1016/j.enbuild.2018.04.052.

[37] K. Cebrat, Łukasz Nowak, Revealing the relationships between the energy parameters of single-family buildings with the use of self-organizing maps, Energy and Buildings 178 (2018) 61 – 70. doi:10.1016/j.enbuild.2018.08.028.

[38] C. Zhang, L. Cao, A. Romagnoli, On the feature engineering of building energy data mining, Sustainable Cities and Society 39 (2018) 508 – 518. doi:10.1016/j.scs.2018.02.016.

[39] M. W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption, Energy and Buildings 147 (2017) 77 – 89. doi:10.1016/j.enbuild.2017.04.038.

[40] H. Saha, A. R. Florita, G. P. Henze, S. Sarkar, Occupancy sensing in buildings: A review of data analytics approaches, Energy and Buildings 188-189 (2019) 278 – 285. doi:10.1016/j.enbuild.2019.02.030.

[41] S. D'Oca, T. Hong, J. Langevin, The human dimensions of energy use in buildings: A review, Renewable and Sustainable Energy Reviews 81 (2018) 731 – 742. doi:10.1016/j.rser.2017.08.019.

[42] C. Sun, R. Zhang, S. Sharples, Y. Han, H. Zhang, Thermal comfort, occupant control behaviour and performance gap – a study of office buildings in north-east China using data mining, Building and Environment 149 (2019) 305 – 321. doi:10.1016/j.buildenv.2018.12.036.

[43] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework, Energy and Buildings 88 (2015) 395 – 408. doi:10.1016/j.enbuild.2014.11.065.

[44] A. Capozzoli, M. S. Piscitelli, A. Gorrino, I. Ballarini, V. Corrado, Data analytics for occupancy pattern learning to reduce the energy consumption of hvac systems in office buildings, Sustainable Cities and Society 35 (2017) 191 – 208. doi:10.1016/j.scs.2017.07.016.

[45] S. Wolf, J. K. Møller, M. A. Bitsch, J. Krogstie, H. Madsen, A markov switching model for building occupant activity estimation, Energy and Buildings 183 (2019) 672 – 683. doi:10.1016/j.enbuild.2018.11.041.

[46] P. Geyer, A. Schlueter, S. Cisar, Application of clustering for the development of retrofit strategies for large building stocks, Advanced Engineering Informatics 31 (2017) 32 – 47. doi:10.1016/j.aei.2016.02.001.

[47] H. Son, C. Kim, Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches, Journal of Cleaner Production 109 (2015) 144 – 151. doi:10.1016/j.jclepro.2014.08.071, special Issue: Toward a Regenerative Sustainability Paradigm for the Built Environment: from vision to reality.

[48] A. Capozzoli, D. Grassi, M. S. Piscitelli, G. Serale, Discovering knowledge from a residential building stock through data mining analysis for engineering sustainability, Energy Procedia 83 (2015) 370 – 379. doi:10.1016/j.egypro.2015.12.212.

[49] M. Jun, J. C. Cheng, Selection of target leed credits based on project information and climatic factors using data mining techniques, Advanced Engineering Informatics 32 (2017) 224 – 236. doi:10.1016/j.aei.2017.03.004.

[50] J. Kim, J.-Y. Hyun, W. K. Chong, S. Ariaratnam, Understanding the effects of environmental factors on building energy efficiency designs and credits: Case studies using data mining and real-time data, Journal of Engineering, Design and Technology 15 (2017) 270–285. doi:10.1108/JEDT-12-2015-0082.

[51] K. Mason, S. Grijalva, A review of reinforcement learning for autonomous building energy management (2019) in press.

[52] A. Garrett, J. New, Scalable tuning of building models to hourly data, Energy 84 (2015) 493 – 502. doi:10.1016/j.energy.2015.03.014.

[53] L. Tronchin, M. Manfren, P. A. James, Linking design and operation performance analysis through model calibration: Parametric assessment on a passive house building, Energy 165 (2018) 26 – 40. doi:10.1016/j.energy.2018.09.037.

[54] Y. Peng, J.-R. Lin, J.-P. Zhang, Z.-Z. Hu, A hybrid data mining approach on BIM-based building operation and maintenance, Building and Environment 126 (2017) 483 – 495. doi:10.1016/j.buildenv.2017.09.030.

[55] C. Jin, M. Xu, L. Lin, X. Zhou, Exploring BIM data by graph-based unsupervised learning, in: ICPRAM, 2018.

[56] Y. Liu, Y. Huang, R. Stouffs, Using a data-driven approach to support the design of energy-efficient buildings, ITcon , Special issue ECPPM 2014 - 10th European Conference on Product and Process Modelling 20 (2015) 80–96.

[57] S. Yarmohammadi, R. Pourabolghasem, D. Castro-Lacouture, Mining implicit 3d modeling patterns from unstructured temporal BIM log text data, Automation in Construction 81 (2017) 17 – 24. doi:10.1016/j.autcon.2017.04.012.

[58] H. Kim, A. Stumpf, W. Kim, Analysis of an energy efficient building design through data mining approach, Automation in Construction 20 (2011) 37 – 43. doi:10.1016/j.autcon.2010.07.006.

[59] S. Tucker, C. B. de Souza, Thermal simulation outputs: exploring the concept of patterns in design decision-making, Journal of Building Performance Simulation 9 (2016) 30–49. doi:10.1080/19401493.2014.991755.

[60] C. B. de Souza, S. Tucker, Thermal simulation software outputs: a framework to produce meaningful information for design decision-making, Journal of Building Performance Simulation 8 (2015) 57–78. doi:10.1080/19401493.2013.872191.

[61] M. Gajzler, Usefulness of mining methods in knowledge source analysis in the construction industry, Archives of Civil Engineering 62 (2016). doi:10.1515/ace-2015-0056.

[62] M. Gajzler, Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry, Technological and Economic Development of Economy 16 (2010) 219–232. doi:10.1515/ace-2015-0056.

[63] V. Ahmed, Z. Aziz, A. Tezel, Z. Riaz, Challenges and drivers for data mining in the AEC sector, Engineering, Construction and Architectural Management 25 (2018) 1436–1453. doi:10.1108/ECAM-01-2018-0035.

[64] C. Bizer, T. Heath, T. Berners-Lee, Linked data - the story so far, International Journal on Semantic Web and Information Systems 5 (2009) 1–22.

[65] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American 284 (2001) 34–43.

[66] P. Pauwels, S. Zhang, Y.-C. Lee, Semantic web technologies in AEC industry: a literature review, Automation in Construction 73 (2017) 145–165. doi:10.1016/j.autcon.2016.10.003.

[67] J. Beetz, J. van Leeuwen, B. de Vries, An ontology web language notation of the industry foundation classes, in: Proceedings of the 22nd CIB W78 Conference on Information Technology in Construction, 2005, pp. 193–198.

[68] P. Pauwels, W. Terkaj, EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology, Automation in Construction 63 (2016) 100–133. doi:10.1016/j.autcon.2015.12.003.

[69] G. F. Schneider, M. H. Rasmussen, P. Bonsma, J. Oraskari, P. Pauwels, Linked Building Data for Modular Building Information Modelling of a Smart Home, in: eWork and eBusiness in Architecture, Engineering and Construction (ECPPM 2018), CRC Press, Copenhagen, Denmark, 2018, pp. 407–414.

[70] M. H. Rasmussen, P. Pauwels, C. A. Hviid, J. Karlshøj, Proposing a central AEC ontology that allows for domain specific extensions, in: F. Bosché, I. Brilakis, R. Sacks (Eds.), Proceedings of the Joint Conference on Computing in Construction, Heraklion, Crete, Greece, 2017.

doi:10.24928/jc3-2017/0153.

[71] G. F. Schneider, Towards aligning domain ontologies with the Building Topology Ontology, in: 5th Linked Data in Architecture and Construction Workshop, Univeristy of Burgundy, Dijon, France, 2017, doi:10.13140/RG.2.2.21802.52169.

[72] D. Bonino, F. Corno, DogOnt - ontology modeling for intelligent domotic environments, in: Proceedings of the International Semantic Web Conference (ISWC), volume 5318 of *Lecture Notes in Computer Science (LNCS)*, 2008, pp. 790–803. doi:10.1007/978-3-540-88564-1_51.

[73] G. Costa, L. Madrazo, An information system architecture to create building components catalogues using semantic technologies, in: A. Mahdavi, B. Martens, R. Scherer (Eds.), Proceedings of the 10th European Conference on Product and Process Modelling (ECPPM), 2014, pp. 551–557, doi:10.1201/b17396-90.

[74] P. Pauwels, W. Terkaj, T. Krijnen, J. Beetz, Coping with lists in the ifcOWL ontology, in: Proceedings of the 22nd EG-ICE International Workshop, 2015, pp. 113–122.

[75] K. McGlinn, A. Wagner, P. Bonsma, L. McNerney, D. O'Sullivan, Interlinking geospatial and building geometry with existing and developing standards on the web, Automation in Construction 103 (2019) 235–250, doi:10.1016/j.autcon.2018.12.026.

[76] J.-P. Calbimonte, H. Jeung, O. Corcho, K. Aberer, Enabling query technologies for the semantic sensor web, International Journal on Semantic Web and Information Systems 8 (2012) 43–63.

[77] A. Sheth, C. Henson, S. Sahoo, Semantic sensor web, IEEE Internet Computing 12 (2008) 78–83.

[78] J.-P. Calbimonte, O. Corcho, A. J. G. Gray, Enabling ontology-based access to streaming data sources, in: The Semantic Web – ISWC 2010, Springer, Berlin, Heidelberg, 2010, pp. 96–111.

[79] X. Wang, X. Zhang, M. Li, A survey on semantic sensor web: Sensor ontology, mapping and query, International Journal of u- and e- Service, Science and Technology 8 (2015) 325–342.

[80] K. R. Llanes, M. A. Casanova, N. M. Lemus, From sensor data streams to linked streaming data: a survey of main approaches, Journal of Information and Data Management 7 (2016) 130–140.

[81] M. Lefrançois, J. Kalaoja, T. Ghariani, A. Zimmermann, D2.2: The SEAS Knowledge Model, Technical Report, ITEA2 12004 Smart Energy Aware Systems, Brussels, Belgium, 2017.

[82] M. Rasmussen, C. Frausing, C. Hviid, J. Karlshøj, Demo: Integrating building information modeling and sensor observations using semantic web, in: Proceedings of the 9th International Semantic Sensor Networks Workshop co-located with 17th International Semantic Web Conference (ISWC 2018), 2018, pp. 48–55. URL: `http://ceur-ws.org/Vol-2213/`.

[83] E. Curry, J. O'Donnell, E. Corry, S. Hasan, M. Keane, S. O'Riain, Linking building data in the cloud: Integrating cross-domain building data using linked data, Advanced Engineering Informatics 27 (2013) 206–219. doi:10.1016/j.aei.2012.10.003.

[84] G. A. Benndorf, D. Wystrcil, N. Réhault, Energy performance optimization in buildings: A review on semantic interoperability, fault detection, and predictive control, Applied Physics Reviews 5 (2018) 041501. doi:10.1063/1.5053110.

[85] S. Hu, E. Corry, M. Horrigan, C. Hoare, M. D. Reis, J. O'Donnell, Building performance evaluation using openmath and linked data, Energy and Buildings 174 (2018) 484 – 494. doi:10.1016/j.enbuild.2018.07.007.

[86] J. O'Donnell, E. Corry, S. Hasan, M. Keane, E. Curry, Building performance optimization using cross-domain scenario modeling, linked data, and complex event processing, Building and Environment 62 (2013) 102 – 111. doi:10.1016/j.buildenv.2013.01.019.

[87] K. McGlinn, B. Yuce, H. Wicaksono, S. Howell, Y. Rezgui, Usability evaluation of a web-based tool for supporting holistic building energy management, Automation in Construction 84 (2017) 154 – 165. doi:10.1016/j.autcon.2017.08.033.

[88] B. Zhong, C. Gan, H. Luo, X. Xing, Ontology-based framework for building environmental monitoring and compliance checking under BIM environment, Building and Environment 141 (2018) 127 – 142. doi:10.1016/j.buildenv.2018.05.046.

[89] J. J. V. Díaz, M. R. Wilby, A. B. R. González, J. G. Munoz, Eeont: An ontological model for a unified representation of energy efficiency in buildings, Energy and Buildings 60 (2013) 20 – 27. doi:10.1016/j.enbuild.2013.01.012.

[90] I. Esnaola-Gonzalez, J. Bermudez, I. Fernandez, A. Arnaiz, Semantic prediction assistant approach applied to energy efficiency in tertiary buildings, Semantic Web 9 (2018) 735–762. doi:10.3233/SW-180296.

[91] R. Hoehndorf, N. Queralt-Rosinach, Data science and symbolic AI: Synergies, challenges and opportunities, Data Science 1 (2017) 27–38. doi:10.3233/ds-170004.

[92] P. Patel, E. Keogh, J. Lin, S. Lonardi, Mining motifs in massive time series databases, in: 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 2002, pp. 370–377. doi:10.1109/ICDM.2002.1183925.

[93] P. Weiner, Linear pattern matching algorithms, in: 14th Annual Symposium on Switching and Automata Theory (SWAT 1973), 1973, pp. 1–11. doi:10.1109/SWAT.1973.13.

[94] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Mining and Knowledge Discovery 8 (2004) 53–87. doi:10.1023/B:DAMI.0000005258.31418.83.

[95] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Mining and Knowledge Discovery 15 (2007) 107–144.

Appendix D. Paper IV

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018). From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches. In: J. Karlshøj, & R. Scherer (Eds.) *eWork and eBusiness in Architecture, Engineering and Construction: Proceedings of the 12th European Conference on Product and Process Modelling (ECPPM)*, Copenhagen, Denmark, pp. 391- 399. London: CRC Press/Balkema.

https://doi.org/10.1201/9780429506215-49

From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches

E. Petrova, K. Svidt & R.L. Jensen
Aalborg University, Aalborg, Denmark

P. Pauwels
Ghent University, Ghent, Belgium

ABSTRACT: Decision-making in design and engineering relies little on knowledge discovered in previous projects and embedded in digital data. Applying analytical computational techniques to available data and processes can be of significant influence for infusing decision-making with the evidence-based character that it is currently lacking. The design environment is where decisions are implemented, therefore, we aim to endow it with knowledge discovered in previous projects and existing buildings. We use an approach that combines data mining and semantic modelling for case-based design (CBD). We investigate the character of the active design environment, what queries can be constructed automatically from the data available in that environment, and how they can be executed against a repository of design models and performance patterns obtained using Knowledge Discovery in Databases (KDD) and various machine learning approaches. We demonstrate this approach on a use case, highlighting its potential for evidence-based design decision support.

1 INTRODUCTION

The advancements in predictive analytics and simulations have led to the implementation of innovative performance assessment models in the building design domain. Yet, many of the decisions taken rely on design assumptions and previous experience, rather than documented evidence. The Architecture, Engineering and Construction (AEC) industry is more information-intensive than ever and that by itself unveils an unprecedented opportunity for discovery of hidden knowledge in the significant heterogeneous datasets generated during the design, construction, and operation of buildings (Soibelman & Kim, 2002; Bilal et al, 2016). Powerful cross-domain techniques such as machine learning and semantic query techniques have made prediction of performance outcomes and knowledge discovery not just possible, but much more accurate and reusable.

Being applied to available data, such approaches carry a powerful potential and can be of fundamental influence to the decision-making process by giving it an evidence-based character (Hamilton & Watkins, 2009). Relevant data sources may include operational building data from sensor networks, Building Information Models (BIM), design brief databases, performance targets relative to the sustainability criteria, etc. By employing the powerful potential of Knowledge Discovery in Databases (KDD) (Fayyad et al, 1996), data mining (Hand et al, 2001) and pattern recognition (Bishop, 2006), evidence can be found in patterns and potentially occurring links between patterns discovered in the data. And while traditional analytical and prescriptive approaches present issues when it comes to high-performance design, a combination of holistic performance-oriented approaches and computational technologies can more effectively contribute to achieving evidence-based decision-making. Besides the available data and the patterns discovered in the data, a decision support system is also essentially influenced by the design development environment, as it is the place that drives queries to any of the knowledge sources that are potentially available.

In this regard, we look specifically at the target data, and how discovered patterns in building operation can be retrieved and used to support the decision-making in new design processes. Therefore, this research effort focuses on enhancing sustainable building design through analytical computational approaches applied in the early design phase. We start from a design environment that is empowered by BIM tools. Furthermore, design brief requirements are considered to be an integral part of the design environment as well. Hence, a Common Data Environment (CDE) takes a prominent place in this research, as the CDE functions as the environment in which all design data is available. From this environment, knowledge is sought for in a pattern retrieval repository, which is based on an open repository of Industry

Foundation Classes (IFC) models collected from previously executed building designs, for some of which motifs (frequent repetitive patterns) and association rules have been discovered.

In this article, we first look into related works (Section 2) aimed at informing building design with knowledge from existing buildings and/or similar designs. In Section 3, we explore the structure of design environments and propose the way in which such systems may be enhanced with evidence-based decision support. Section 4 documents the performed experiment, which consists of (1) a data repository containing building semantics and performance data, (2) a specifically considered building design, and (3) the tests conducted towards matching them. Section 5 discusses the results and future works, thereby leading to Section 6, which concludes this article.

2 RELATED WORKS

Using KDD and data mining approaches in AEC has gained momentum with regards to improvement of building performance. Promising advancements lie within the use of machine learning for model predictive control (Drgona, 2018), metamodelling for design space exploration (Geyer and Schlueter, 2014; Østergaard et al., 2018), use of data analytics for improvement of energy performance and building occupancy (Ahmed et al., 2011; Fan et al., 2015a), etc. Most prominently, research has shown great advancements related to use of data analytics for improvement of facility management and building operation. Included here are anomaly and fault detection diagnostics in systems operation, extraction of energy use and occupant behaviour patterns, improvement of occupant comfort, etc. (Fan et al., 2015b; Fan et al., 2018).

With regards to the use of KDD for design decision support, research efforts include pattern recognition in simulation data and extraction of information from BIM design log files (Yarmohammadi et al., 2016), extraction of 3D modelling patterns from unstructured temporal BIM log text data (Yarmohammadi et al., 2017), use of data-driven approaches to design energy-efficient buildings by mining of BIM data (Liu et al., 2015) and use of simulation data mining for energy efficient building design (Kim et al., 2011). Reuse of similarities in decision support has also been widely recognised in design practice. This is prominently present in case-based reasoning (CBR), which provides decision makers with a problem solving framework involving recalling and reusing previous knowledge and experience (Aamodt and Plaza, 1994). The use of CBR in design practice (case-based design (CBD)) differs with regards to the method of implementation. For instance, Dave et al. (1994) present a design system enabling case adaptation and combination for a more efficient generation of new design cases. Both Heylighen and Neuckermans (2000) and Richter et al. (2007) demonstrate the implementation of CBD in architecture to support knowledge renewal and exchange between designers. Eilouti (2009) further explores the possibility for recycling architectural design knowledge by reuse of design precedents.

In the context of sustainable building design, Xiao et al. (2017) develop an experience mining model for solving green building design problems by CBR, and thereby assist the decision maker in finding solutions. Shen at al. (2017) introduce an integrated system of text mining and CBR for retrieval of similar green building cases when producing new green building designs. In terms of energy efficiency, Abaza (2008) presents a model, where the computer evaluates design alternatives suggested by the designer and generates a matrix of design solutions. More recent approaches include that of Sabri et al. (2017) who apply CBR and graph matching techniques for retrieval of similar architectural floor plans in the early design stages. Ayzenshtadt et al. (2016) investigate the potential of rule-based and case-based retrieval coordination for architectural design search. Weber et al. (2010) propose a sketch-based retrieval system based on CBR and shape detection technologies, which gives access to a semantic floorplan repository. These approaches typically capture semantics in topology graphs, which is less complex and detailed compared to the rich semantics of BIM data.

However, despite coming a step closer to realizing the targeted process, these efforts rely mostly on design patterns for improvement of the design, or use performance patterns for improvement of building and system operation. Using knowledge discovered in performance data to influence design decision-making and improve future building design processes is an area that is rarely explored in detail. Furthermore, the combined use of semantics, KDD, and CBD is seldom achieved. Therefore, in this article, we aim to combine these three approaches for influencing design decisions using both design and operational building data.

3 PROPOSED TECHNICAL APPROACH

The way in which design professionals approach decision-making is characterized by iterative problem-solution cycles, in which solutions are widely based on tacit knowledge. Each design iteration explores a problem/solution space, which leads to a repetitive co-evolution of problems and solutions (Dorst and Cross, 2001). Figure 1 depicts that process, during which the design team aims to converge in the problem and solution spaces.

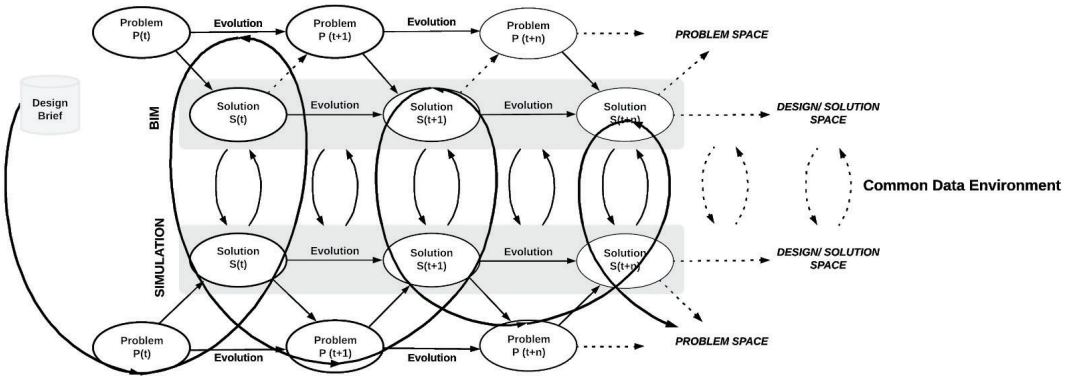

Figure 1. Problem-Solution iterations in collaborative design.

Convergence brings the team closer to a solution that fulfills the design brief and the performance targets, while avoiding widening of the cycles.

A typical design environment may include BIM authoring tools, parametric design tools, simulation tools, etc., by the use of which design professionals iterate through a number of proposals, both individually and in a collaborative manner. The generated design data is stored in the CDE. To be able to influence the above process, performance data and knowledge discovered in data need to be presented to the decision maker in the form of useful design alternatives matching the stated objectives. We therefore aim to connect the active design environment with a repository that collects data available from previous projects and the corresponding existing buildings. The data in the repository has various heterogeneous origins, representations, and purposes. Knowledge Discovery can be applied to this data, thereby following the KDD process defined by Fayyad et al. (1996), which consists of five steps. They include selection, cleansing, transformation, mining, and interpretation / evaluation of the data. It is important to note that a significant part of the workload is dedicated to data selection, cleansing and transformation. Furthermore, the evaluation step is critical to the interpretation of the meaning of the patterns found in the data. This study follows these five steps in creating the repository of design data with associated discovered patterns.

In this study, we aim to connect the outlined repository with the active design environment. This can be any BIM tool or the CDE itself. Recent initiatives aim at making the data available in an integrated manner using web technologies, both in the context of BIM tools and the CDE. In this regard, web technologies can enable a web-compliant and data-oriented information management approach. Such an approach is desirable as it (1) allows the integration of heterogeneous data sources, (2) enables federated query techniques over diverse data repositories for advanced information retrieval and (3) provides a well-defined formal data structure to capture building semantics. This results in a design environment as outlined in Fig. 2, with BIM tools on the left, and a web-based CDE on the right.

The adoption of web technologies for representing information in a design environment can be realized using a decentralized graph database approach. Promising in this regard are linked data and semantic web technologies (Berners-Lee et al., 2001; Pauwels et al., 2017a), which allow to build a decentralized web of semantic information, consisting of various repositories with relevant building data. Such repositories can contain various kinds of data, including design brief data, user logs, BIM models, performance data, etc. For the purpose targeted in this paper, we therefore propose a semantic integration layer, which maintains the links between the individual datasets (Fig. 2). The semantic integration layer has a thin and modular structure which captures the semantics of available data, keeping the original data sources in their optimized structures.

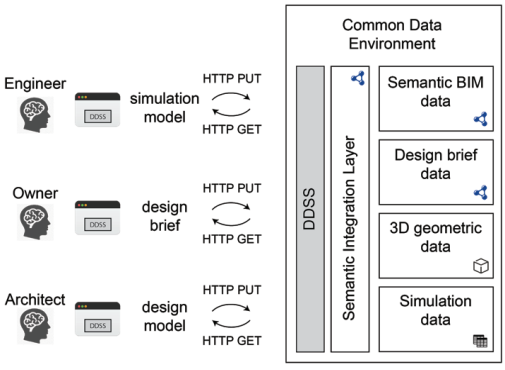

Figure 2. Integration of datasets in a web-based design environment.

4 USE CASE EXPERIMENT

In this section, we test the proposed approach and consider how the active design environment can be connected to a repository of design data that is enriched with patterns obtained using a KDD process, employing motif discovery and association rule mining algorithms (Fu, 2011; Patel et al. 2002). Each part of the experiment is documented here, including the repository (Section 4.1), the active design case (Section 4.2), and matching both (Section 4.3).

4.1 *Building the information retrieval repository*

A rich data repository should include heterogeneous data for multiple diverse buildings. That includes not only building models, but also design briefs, simulation and sensor data, and so forth. In this case, however, we limited to working with a collection of building models in the IFC data model, which was previously set up in the context of the performance benchmark in Pauwels et al. (2017b). At that time, the repository consisted of 369 building models of diverse size, origin and kind. The current version of the repository[1] consists of 531 IFC models.

As a first step, all models have been converted into linked data. This step makes the data easy to query, as linked data technologies come with an out-of-the-box query language (SPARQL), as opposed to the STEP and EXPRESS technology used by IFC. This conversion is done using an open source IFC to Linked Building Data converter[2]. The result is a set of RDF graphs in TTL format that are compliant with the BOT (Rasmussen et al., 2017), PRODUCT, and PROPS ontologies[3]. For this study, the conversion to LBD excludes geometry from the data, leaving only the semantic backbone and product data for the building models. Geometric data may be converted to linked data and made available, but is less useful for the purpose of the current semantic information retrieval effort. In order to be useful for information retrieval, the raw geometry should be processed first to contain semantically useful concepts (e.g. above, below, next to), which is out of scope for the current study.

The final result is a collection of two Stardog triple store databases, with in total 36 Million triples (24.951.647 triples and 11.425.589 triples). The data was spread over two databases, aiming to test and validate a decentralized information structure and a federated query approach. The data includes 372 *bot:Building* instances, 3,523 *bot:Zone* instances, 2,117 *bot:Space* instances, and 615,452 *bot:Element* instances. The *bot:Element* instances also have a more specific product type. For instance, one of the repositories includes 45 distinct product types, including *product:Wall, product:Fastener, mep:FlowTerminal, product:Pile*, etc. Each of these instances has a number of associated properties. Clearly, the majority of available triples consists of properties associated to building elements. At the moment, these properties come in various languages and notations, which makes it difficult to query them. Ideally, they should follow an ontology, which is the purpose of the PROPS ontology[4].

For some of the models in this repository, sensor data is available from the corresponding existing buildings. The sensor data is also modelled using linked data best practices[5]. More particularly, we used the SOSA ontology to describe the relationships between the spaces and the contained sensor nodes (data points), each of which has individual sensors, with observations and results. All data modelling is done according to the SOSA ontology, giving a semantic representation of the sensors and their observations and values in context of the spaces. The data values of the sensor data are not directly included in the semantic graph, in order not to make that graph too complex. Instead, links are maintained to the original locations where the sensor data is stored. This is done using a custom *gig:values* datatype property added to specific sensor nodes. These properties point to a web address that returns the data values as requested using the HTTP protocol. One is able to add attributes to an HTTP request, thereby setting query parameters such as time frame and refresh rate (e.g. from=now-30d&to=now&refresh=30s). The result includes the pointer to the data stream for a sosa:Result of a sosa:Observation. A short example snippet is provided in the Listing below:

```
inst:room_16
  rdf:type bot:Space ;
  gig:hasSensorNode inst:sensorNode_0000014 ;
  gig:spaceType "Cafe" ;
  rdfs:label "Cafe" .

inst:sensorNode_00000014
  rdf:type gig:SensorNode ;
  rdfs:label "00000014" ;
  gig:observation "Indoor climate" ;
  gig:purpose "Thermal comfort in the lobby during big events when there is a gathering of a lot of people." ;
  sosa:hosts inst:sensor_00000014_1 ;
  sosa:hosts inst:sensor_00000014_2 ;
  sosa:hosts inst:sensor_00000014_3 ;
  sosa:hosts inst:sensor_00000014_4 ;
```

[1] http://smartlab1.elis.ugent.be:8889/IFC-repo/
[2] https://github.com/jyrkioraskari/IFCtoLBD
[3] https://www.w3.org/community/lbd/
[4] https://github.com/w3c-lbd-cg/props
[5] https://www.w3.org/TR/ld-bp/

```
    sosa:hosts inst:sensor_00000014_5 ;
    sosa:hosts inst:sensor_00000014_6 ;
    gig:placement "Placed on a column in the cafe
without direct sunlight." .

  inst:sensor_00000014_1 ;
    rdf:type sosa:Sensor ;
    sosa:madeObservation inst:observation_1 ;
    sosa:observes inst:obsProperty_1 ;
    rdfs:label "00000014_1" .

  inst:result_1 rdf:type sosa:Result ;
    rdfs:label "Result of observation of Relative Hu-
midity" ;
    gig:values
    "https://gigantium.dk/Gigantium2018in-
 stances?orgId=1&datastream=true" .
```

To make the use of collected sensor data more effective and based on the stated goal, multiple KDD techniques can be applied. We have specifically tested this approach in this study for some of the available sensor data. In this case, a combination of motif discovery and association rule mining has been applied to time series data. The detailed description and implementation of the KDD steps is performed in advance and is out of scope for the current paper. The resulting motifs and the related co-occurrence rules are added to the graph using a separate in-house developed pattern matching ontology. In more detail, a sensor node in the graph is directly linked to an instance of a *pattern:AssociationRule*, which furthermore links to a left hand side and right hand side in the rule. Both left hand and right hand side concepts furthermore link to *pattern:motif* concepts, such as M1 and M5 (Fig. 3).

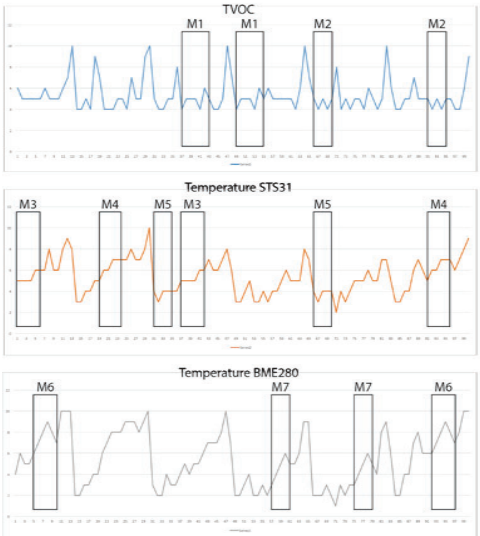

Figure 3. Obtained observation data and discovered patterns.

In this example, these motifs occur in temperature and Total Volatile Organic Compounds (TVOC) observations for a cafeteria in a public building. These motifs are semantically described as well, eventually including the exact data sensor values for those observations.

4.2 *The active design case*

In addition to the repository of design models with sensor data and performance patterns, an active design model was selected, which forms the starting point for knowledge retrieval. We use a design model of a healthcare facility (Fig. 4 and 5), which is a part of the active design environment (in this case Autodesk Revit) and hence can be used to retrieve relevant knowledge from the repository (see Section 3).

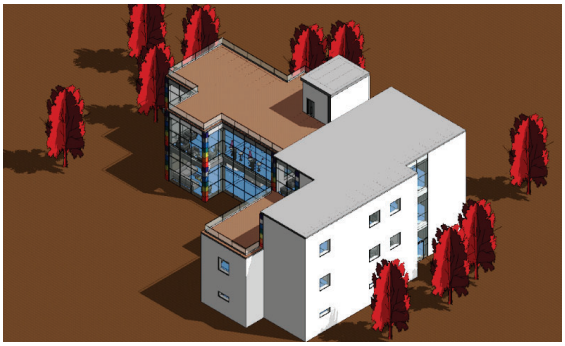
Figure 4. Revit design model of a healthcare facility.

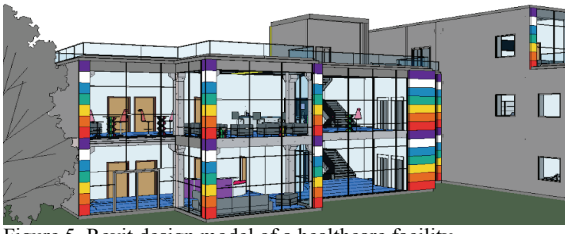
Figure 5. Revit design model of a healthcare facility.

The building design consists of two main parts (one wing with public access and another one accessible by medical professionals and patients only), connected with a connection spine. One part of the building contains the entrance, visitors' lobby, cafeteria and public spaces; the other part contains the patient wards, examination, operating and recovery rooms, staff rooms, etc. The basement area contains all necessary technical and equipment rooms.

In addition to the BIM design model in Revit, a number of design brief requirements are available. As they were unstructured in this case (textual document), we decided to not use them, in contrast to what is often done in related works with CBR and text mining techniques. Instead, we consider in our case a semantic building model directly linked to a semantic

representation of design brief requirements. This semantic data can be used to perform case retrieval in the repository documented before, to then inform the designer of factual performance data in existing buildings.

4.3 *Information retrieval and pattern matching*

In order to obtain reference knowledge from the building data repository, a direct matching needs to be made between the new case and existing cases (cfr. CBR). Such matching can occur in a number of ways. As we have seen, most existing works perform geometric spatial layout matching using topology graphs. Even though many of these topology graphs have some semantics, the available semantics is in this case a lot more complex and rich. The semantics embedded in the design brief and the design model allows to perform semantically more specific queries and thus better matching. This does not rule out topology graph matching. Also user action log data can be useful for retrieving relevant cases. Depending on what actions the users take, their intentions may be tracked in a more intelligent way, thus improving the matches with the building knowledge repository.

As preliminary design decisions are made in an early planning stage that relies heavily on space types and configurations, for this case study we focus on matching cases based on space type. Obviously, a full implementation can take into account a lot more of the available semantic data, aiming to match system configurations, material choices, expected usage patterns, and so on.

Matching the active design model is thus implemented using SPARQL queries, such as the one listed in Listing 2. This query shows how the repository is queried for buildings with spaces of type “cafeteria”, aiming to retrieve not only those buildings, but also the corresponding performance data and patterns obtained using the data mining techniques briefly mentioned in Section 4.1. Querying is done through a federated query approach. The two repositories that are built for this use case are queried using the SERVICE keyword, as indicated in the Listing below:

```
SELECT ?b ?s ?o
WHERE {
SERVICE
<http://localhost:5820/BuildingRepo/query> {
        ?b rdf:type bot:Building .
        ?b bot:hasSpace ?s .
        ?s rdf:type bot:Space .
   ?s props:categoryDescription “cafeteria” }
SERVICE
<http://localhost:5820/BuildingRepo1/query> {
        ?b rdf:type bot:Building .
        ?b bot:hasSpace ?s .
        ?s rdf:type bot:Space .
   ?s props:categoryDescription “cafeteria” }
}
```

These queries can be implemented in a plug-in for the corresponding design environment, or directly from the CDE, in which case more alternative data is available (briefs, logs, simulation data). The returned Unique Resource Identifiers (URIs) for spaces and buildings provide reference points for obtaining more data. These URIs can be used by plugins or CDE to subsequently query for building performance patterns that are available for the retrieved buildings and spaces.

In our case, the query in Listing 2 returns, among others, a cafe that is part of a visitors’ lobby in a sports and cultural centre, for which operational data and performance patterns are available (Fig. 3). This data can be directly provided to the end user. Hence, users can be provided not only with a link to sample existing buildings of the kind they are developing (in this case, the bar in the hospital), they can also retrieve the knowledge about that place which is captured in patterns obtained from a KDD process. Being able to obtain this information during a design process is considered of utmost relevance in informing design decisions.

5 RESULTS AND DISCUSSION

The presented use case with the data repository consisting of 531 models and the healthcare facility design model provides a useful context to evaluate the proposal for decision support using a combination of CBR, KDD, and semantics from within a BIM environment. Results and discussion thus focus on those three main topics.

First and foremost, CBR provides a useful theoretical background for the given proposal around design decision support. In order to be fully effective, it would be useful to extend the amount and diversity of the data that is used, both to document the cases in the repository and to inform queries. In this regard, the availability of a CDE with user log data, design requirements, and performance data is potentially of tremendous relevance.

Second, the semantics provide effective and rich means to retrieve relevant cases. The semantic richness provides great opportunities to outperform case retrieval using topology graphs and text mining approaches. Nevertheless, there are also some boundaries. Namely, the effectiveness of the system relies a lot on the expressiveness and formal rigor of the ontologies used for capturing semantics. In this case, the *props:categoryDescription* predicate was used, for example, to retrieve spaces of a particular type; yet, very different predicates are used as well, making it difficult to cover an entire dataset. Also, the diversity of languages in a dataset is difficult to cope with. In

this regard, a data dictionary that provides translations between terms is of important relevance.

Finally, one of the most important parts is the KDD process involved in retrieving the performance patterns and associations between them. The KDD process itself is out of scope in this article, yet, the results of that process are directly embedded in the knowledge graph. On-demand data mining is thus not performed. Such on-demand data mining, as well as the actual interpretation/evaluation of the discovered patterns is essential to turning the results into actionable knowledge. Therefore, user-driven KDD may be of relevance to be considered in future research.

6 CONCLUSION

In this article, we look into the ways in which knowledge about existing buildings and their performance patterns can be made accessible in an active design environment to give design processes a more evidence-based character. We particularly investigate how existing CBR approaches can be improved using a combination of BIM, KDD, and semantic data modelling, thereby aiming to enable BIM-based information retrieval in support of sustainable design. The article presents a technical approach, which indicates how decision support can be embedded in a BIM-based design environment and common data environment (CDE). The proposed technical approach is tested in a case study environment consisting of a building data repository and active design model of a healthcare facility. The building data repository consists of 531 building models, for some of which sensor data is available. All data is represented in semantic graphs and made available in a triple store using latest developments and techniques in linked data best practices. Data mining is performed over the sensor data using motif discovery and association rule mining. Finally, a number of semantic queries show how cases can be retrieved that match the active design model, including the retrieval of performance patterns. As such, the potential of the proposed technical approach is demonstrated for case retrieval in support of evidence-based sustainable design.

REFERENCES

Aamodt, A., Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. IOS Press, Vol. 7: 1, 39-59.

Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H., Menzel, K. (2011). Mining building performance data for energy efficient operation. Advanced Engineering Informatics 25, 341–354.

Ayzenshtadt, V., Langenhan, C., Roth, J., Bukhari, S. S., Althoff, K.-D., Petzold, F., and Dengel, A. (2016). Comparative evaluation of rule-based and case-based retrieval coordination for search of architectural building designs. 24th International Conference on Case Based Reasoning, Atlanta, GA, USA. Springer, Berlin, Heidelberg.

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web, Scientific American, pp. 29-37.

Bilal, M., Oyedele, L., Qadir. J., Munir, K., Ajayi, S., Akinade, O., Owolabi, H., Alaka, H., Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced Engineering Informatics 30, 500–521.

Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.

Dave, B., Schmitt, G., Faltings, B., Smith, I. (1994). Case based design in architecture. Artificial Intelligence in Design- AID '94, J. Gero and F. Sudweeks (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, 145-162.

Dorst, K. & Cross, N. (2001). Creativity in the design process: coevolution of problem-solution. Design Studies 22(5), 425-437.

Drgoňa, J., Picard, D., Kvasnica, M., Helsen, L. (2018). Approximate model predictive building control via machine learning. Applied Energy 218, 199-216.

Elouti, B.H. (2009). Design knowledge recycling using precedent-based analysis and synthesis models. Design Studies, 30, 340-368.

Fan, C., Xiao, F., Madsen, H. & Wang, D., (2015a). Temporal knowledge discovery in big BAS data for building energy management. Energy and Buildings, Vol 109, pp. 7589.

Fan, C., Xiao, F. & Yan, C., (2015b). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automation in Construction, Vol 50, pp. 8190.

Fan, C., Xiao, F., Li, Z., Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings, 159, 296–308 .

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 37-54.

Fu, T.C. (2011). A review on time series data mining, Engineering Applications of Artificial Intelligence, 17, 164–181.

Geyer, P. and Schlueter, A. (2014). Automated metamodel generation for Design Space Exploration and decision-making–A novel method supporting performance-oriented building design and retrofitting. Applied Energy, 119, 537-556.

Hand D., Mannila H., Smyth P. (2001). Principles of Data Mining. MIT Press, Cambridge.

Hamilton, D.K. and Watkins, D. (2009). Evidence-Based Design for Multiple Building Types. John Wiley & Sons, New Jersey, USA.

Heylighen, A., Neuckermans, H. (2000). DYNAMO: A Dynamic Architectural Memory On-line. Educational Technology & Society (3)2.

Kim, H., Stumpf, A., Kim, W. (2011). Analysis of an energy efficient building design through data mining approach. Automation in Construction, 20, 37–43.

Liu, Y., Huang, Y.C., Stouffs, R. (2015). Using a data-driven approach to support the design of energy-efficient buildings. Journal of Information Technology in Construction, 20, 80-96 .

Østergård, T., Jensen, R.L., Maagaard, S.E. (2018). A comparison of six metamodeling techniques applied to building performance simulations. Applied Energy, 211, 89-103.

Patel, P., Keogh, E., Lin, J., Lonardi, S. (2002). Mining Motifs in Massive Time Series Databases. In proceedings of the 2002 IEEE International Conference on Data Mining.

Pauwels, P., Zhang, S. & Lee, Y.C. (2017a). Semantic web technologies in AEC industry: A literature overview. Automation in Construction 73, 145-165.

Pauwels, P., de Farias, T.M., Zhang, C., Roxin, A, Beetz, J., De Roo, J., Nicolle, C. (2017b). A performance benchmark over semantic rule checking approaches in construction industry. Advanced Engineering Informatics 33, 68-88.

Rasmussen, M.H., Pauwels, P., Hviid, C.A. & Karlshøj, J. (2017). Proposing a central AEC ontology that allows for domain specific extensions. Proceedings of the Joint Conference on Computing in Construction (JC3), 237-244.

Richter, K., Heylighen, A., and Donath, D. (2007). Looking back to the future-an updated case base of case-based design tools for architecture. Knowledge Modelling eCAADe, 25, 285–292.

Sabri, Q.U., Bayer, J., Ayzenshtadt, V., Bukhari, S.S., Althoff, K.D., Dengel, A. (2017). Semantic Pattern-based Retrieval of Architectural Floor Plans with Case-based and Graph-based Searching Techniques and their Evaluation and Visualization. In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, 50-60.

Shen, L., Yan, H., Fan, H., Wu, Y., Zhang, Y. (2017). An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. Building and Environment, 124, 388-401.

Soibelman, L. and Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. Journal of Computing in Civil Engineering, 16(1), 39–48.

Weber M., Langenhan C., Roth-Berghofer T., Liwicki M., Dengel A., Petzold F. (2010) a.SCatch: Semantic Structure for Architectural Floor Plan Retrieval. Case-Based Reasoning. Research and Development. Lecture Notes in Computer Science, vol 6176. Springer, Berlin, Heidelberg.

Xiao, X., Skitmore, M., Hu, X. (2017). Case-based reasoning and text mining for green building decision making. Energy Procedia, 111, 417 – 425.

Yarmohammadi, S., Pourabolghasem, R., Shirazi, A., Ashuri, B. (2016). A sequential pattern mining approach to extract information from BIM design log files. 33rd International Symposium on Automation and Robotics in Construction., 174-181.

Yarmohammadi, S., Pourabolghasem, R., Castro-Lacouture, D. (2017). Mining implicit 3D modeling patterns from unstructured temporal BIM log text data. Automation in Construction, 81, 17–24.

Appendix E. Paper V

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019, under review). Crowdsourcing building performance patterns for evidence-based decision support in sustainable building design. Submitted to *Automation in Construction*.

Crowdsourcing Building Performance Patterns for Evidence-Based Decision Support in Sustainable Building Design

Ekaterina Petrova[a,*], Pieter Pauwels[b], Kjeld Svidt[a], Rasmus Lund Jensen[a]

[a]*Department of Civil Engineering, Aalborg University, Aalborg, Denmark*
[b]*Department of Architecture and Urban Planning, Ghent University, Ghent, Belgium*

Abstract

The advancements in Building Information Modelling (BIM), Building Monitoring Systems (BMS) and machine learning have made the discovery of hidden insights and performance patterns in operational building data possible and highly accurate. Semantic web technologies play a fundamental role in terms of knowledge representation and also enable the reuse of the discovered insights. Such knowledge can be of particular significance for decision-making support in sustainable BIM-based design. However, this requires patterns discovered with traditional data mining techniques to be attributed with semantics, so that they can be machine-interpretable and reusable in BIM-based workflows. Therefore, this article investigates how semantic data modelling and crowdsourcing techniques can contribute to the semantic enrichment of motifs and association rules discovered in indoor environmental quality data. Using crowdsourcing techniques for interpretation of building performance patterns by domain experts allows to build distributed knowledge graphs of building data, enriched with contextualised operational performance knowledge. That enables both analyses that are not achievable only with traditional data mining techniques, as well as their reuse in an evidence-based building design setting. The article presents a proof of concept for a crowdsourcing mechanism that allows to attribute meaning to building performance patterns through semantic annotation and classification. We elaborate on the results and discuss the potential that distributed linked building data graphs enriched with patterns and annotated using crowdsourcing techniques have for design decision support. The article outlines the technical barriers that need to be overcome to fully implement the suggested system for adoption in real environments and BIM-based workflows.

Keywords: BIM, linked data, building performance, sustainable design, decision support, crowdsourcing, data mining

1. Introduction

The improvement of building performance is a crucial target, considering the significant contribution of the built environment to the global energy consumption, carbon footprint and environmental deterioration. The advent of powerful computational paradigms within and beyond Architecture, Engineering and Construction (AEC) has unlocked great potential when it comes to the use of advanced analytics to improve building design decision-making, and, by effect, building performance itself. As a result, numerous research efforts aim to utilise the technological advancements within machine learning, semantic web technologies, simulation and modeling, to improve building performance. A main driver in terms of digitalisation in AEC has been Building Information Modelling (BIM), which has redefined the performance-oriented integrated workflows in building design and engineering practice [1, 2, 3].

In that relation, semantic data modelling [4] (symbolic AI) and pattern recognition [5, 6] (statistical AI) have also established themselves as essential, complementary to BIM technologies in the shift towards digitalisation in the industry. To-

*Corresponding author
Email addresses: ep@aau.civil.dk (Ekaterina Petrova), pipauwel.pauwels@ugent.be (Pieter Pauwels), ks@aau.civil.dk (Kjeld Svidt), rlj@aau.civil.dk (Rasmus Lund Jensen)

Preprint submitted to Automation in Construction

gether, these computational paradigms can assist advanced simulation-based approaches for building performance enhancement. A number of research initiatives have investigated the adoption of these techniques in the AEC industry. Most of the existing efforts hereby focus either on the statistical side of AI (machine learning) for knowledge discovery in operational building data or BIM data *or*, or on the symbolic AI side (semantics) for ontology engineering, representation and web-based exchange of building data. Researchers in the former area aim to build decision support systems based on self-learning or expert-taught machines, whereas researchers in the latter area aim at the structured definition of data and engineering of information exchange mechanisms.

Most important for the future of the AEC industry, and for this article, is the appropriate combination of statistical and symbolic AI methods in support of the AEC stakeholder. Machine learning algorithms are powerful when it comes to discovery of hidden insights in data, but without being interpreted, these insights are merely analytical output with no ability to influence decision-making in a structured way. (Section 2) will hereby outline the main concepts and contributions in these areas in terms of building performance improvement and design decision support.

Therefore, this study aims to enrich motifs and association rules discovered in operational building data with indoor environmental expertise and store the results in a full semantic graph of the corresponding building where data originates from. This will transform the patterns from exploratory statistical analysis output into machine-readable semantic data attributed with domain expertise, thereby making them applicable in evidence-based design decision support. To achieve this goal, we aim to crowdsource building performance patterns and engage domain experts directly, by allowing them to annotate and interpret these patterns relatively to the source, environment, as well as various static and dynamic parameters essential to the early design stages.

In this regard, this article builds on initial studies, which demonstrate the opportunities that machine learning and semantic web technologies provide in terms of knowledge discovery in indoor environmental quality sensor observations and semantic representation of the discovered performance patterns for evidence-based design decision support [7, 8, 9]. The current article briefly outlines the methodology used for knowledge discovery and representation defined in these studies, after which we use a new use case and discovered association rules to demonstrate the crowdsourcing system for contextualisation and semantic annotation of the discovered knowledge before storing it in the performance-enriched semantic building graph.

The article starts by presenting a review of the most relevant knowledge engineering and crowdsourcing practices (Section 2). We then build on previously documented efforts applying Knowledge Discovery in Databases (KDD) for pattern retrieval from operational building data, including storing the patterns together with the actual building data in a semantic building graph. The meaning of KDD results is of utmost importance to decision-making, but both their interpretation and implementation in design is not straightforward. Thus, in this work we set out a method that allows to disambiguate the discovered knowledge by using domain expertise, semantic data modelling and crowdsourcing techniques (Section 3). We then outline the implementation of this method and demonstrate it with a use case (Section 4). Finally, the last sections discuss the results, present final remarks and outline future work (Section 5 to 6).

2. Knowledge engineering in performance-oriented design

2.1. Knowledge Discovery in Databases

Fayyad et al. [5] define KDD as an overall process, in which knowledge is an end product of a data-driven discovery. Data mining is a step in that process, which relies on dedicated algorithms to discover regularities or irregularities in the data according to a defined goal. The authors define five main steps in the KDD process, i.e. selection, preprocessing, transformation, data mining and interpretation / evaluation [5]. In that context, Hand et al. [10] in turn extend the definition of data mining as *"the analysis of large observational datasets to find unsuspected relationships and summarise the data in novel ways so that data owners can fully understand and make use of the data"*. Fayyad et al. also summarise six main data mining categories,i.e. classification, clustering, association rule mining, regression, summarisation and anomaly detection. Han et al. [11] divide these into two general types of approaches: predictive and descriptive. With

regards to the data source, Lausch et al. [12] differentiate between numeric and categorical data, text, web, media, time series and spatial data mining.

With regards to the types of building data and the knowledge discovery goal, Petrova et al. [7] provide an extensive definition of KDD approaches according to the different building data types (numeric data, semantic BIM data, geometric data, sensor data, etc). Because of the abundance of spatio-temporal data, the AEC industry can benefit from mining time series data and spatial data. Shekhar et al. [13] indicate that extracting interesting patterns and associations from such complex data with plenty of dependencies and spatio-temporal correlations is more difficult than mining traditional numeric and categorical data. Finally, spatio-temporal data mining can target sources that provide not only spatial data, but also temporal data, for example in the case where spatial data is augmented with time series data from diverse sensors in buildings or infrastructure.

A significant body of literature has investigated the use of data mining for building energy management and performance enhancement in the last decade. As data mining is not the main focal point of this research effort, an extensive literature review on the topic is not presented here, but the main applications for energy efficiency and sustainable building design usually relate to prediction of energy use and demand [14], predictions related to occupant behaviour [15], fault detection and diagnostics for building systems [16], optimal modelling and control strategies [17], extracting and explaining energy consumption patterns [18]. Other researchers have investigated the use of semantic data modelling, neural networks and data mining for building energy management [19], etc. As can be seen from these examples, the use of KDD is usually related to improvement of the operational building performance. Using the discovered knowledge to improve future building design processes has not been investigated in such detail in research. Examples of mining BIM data and simulation data for extraction of useful patterns in building design can be seen in Yarmohammadi et al. [20], but these efforts do not consider measured performance data.

2.2. Semantic data modelling

Further to the progress made in the use of KDD for improving the performance of the built environment, a lot of progress has been made in knowledge formalisation and data exchange using semantic web technologies. From a web of documents, the World Wide Web has evolved into a 'Web of Data' (Linked Open Data cloud) [21]. The term Linked Data was first defined by Tim Berners-Lee in 2006[1] and has now enabled world-wide publication of 5-star open data[2]. This implies defining data according to the Resource Description Framework (RDF)[3] data model and interlinking it with other RDF-based datasets on the web. The Web of Data relies on formal vocabularies or ontologies so that data can easily be used in combination with query and rule languages (e.g. SPARQL[4], SHACL[5], SWRL[6], RIF[7], and so forth). Ontologies can be defined using RDFS[8] and OWL[9] and give 'meaning' to the data, thereby contributing significantly to the Semantic Web as conceived by Berners-Lee et al. [4].

Due to their potential for distributed knowledge formalisation on a global level, linked data and semantic web technologies have received major attention in the AEC industry. A comprehensive overview on the application of semantic web technologies in the AEC industry is documented by Pauwels et al. [22]. Among the most notable initiatives is the transformation of the Industry Foundation Classes (IFC) into an OWL ontology (ifcOWL) [23]. The ifcOWL ontology was built to match the original EXPRESS schema as closely as possible, thus allowing a round-trip conversion (lossless conversion). However, this has lead to a very large ontology, which highly resembles the IFC schema, i.e. very difficult to extend, complex, and not modular. This has started several research initiatives that aim to define ontologies for Linked Building Data (LBD), which do not rely that strongly on the IFC data model, yet cover similar concepts.

At the moment, an ecosystem of modular domain ontologies is available, each covering parts of what can also be exchanged with IFC (Fig. 1 and 2). In principle, a small central ontology captures terms as 'Building', 'Space', 'Element' and takes a central role. As of Rasmussen et al. [24], standardisation of these terms is aimed at within the W3C LBD

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://5stardata.info/
[3] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
[4] https://www.w3.org/TR/rdf-sparql-query/
[5] https://www.w3.org/TR/shacl/
[6] https://www.w3.org/Submission/SWRL/
[7] https://www.w3.org/TR/rif-overview/
[8] https://www.w3.org/TR/rdf-schema/
[9] http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

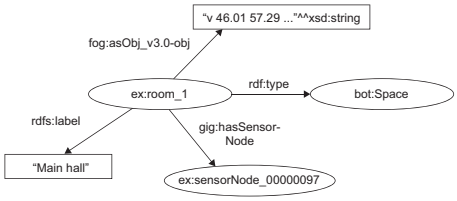

Figure 1: An example LBD graph.

Community Group[10] in the form of a central Building Topology Ontology (BOT)[11]. Starting from this central BOT ontology, alignments can be made with other domain ontologies [25]. As a result, the industry can lean on a modular set of ontologies [26], yet still rely on a stable standard at the core. Besides topology, other modules in the W3C LBD CG focus on products, properties, and geometry [27].

2.3. Crowdsourcing for retrieval of domain expertise and interpretation of patterns

Semantic modelling and data mining allow not only pattern discovery and storage, but also attribution with semantic annotations capturing domain expertise. For example, motif discovery and association rule mining in operational data were demonstrated in Petrova et al. [9], resulting in a semantic graph including performance patterns. Furthermore, it was investigated how such patterns may be retrieved [8] by the end users in a design team. As such, an appropriate combination of KDD and semantic data modelling has already been achieved. However, the interpretation of the motifs and association rules in terms of indoor climate and building performance, and why they emerge, has not been explored in detail, and is the objective of this article. Important to note here is that the focus of this study is to provide the necessary infrastructure for such an interpretation, which would allow not only capturing of domain expertise, but also its retrieval and reuse. An in-depth analysis of the performance patterns and their precise meaning is therefore out of scope in this article.

Instead of storing the patterns and the reasons for their emergence in one "single-opinion, always true" semantic graph, we aim to include domain experts in a continuous evaluation. This would not only improve the underlying knowledge base of the suggested system (improving pattern recognition using a feedback loop), but would also directly engage experts and design teams (users) with performance patterns and their interpretations. Further scientific innovation lies in applying pattern recognition to the resulting patterns. This may result in patterns of patterns or clusters of patterns, which may, in turn, be used for expert decision support within BIM tools aiming at particular buildings or conditions.

A number of techniques are generally available for the retrieval of domain expertise and for interpretation of patterns. In a semantic web environment, the semantic richness of data is key. Hence, a lot of focus has always been put on the ontology engineering part of the semantic web domain. The ontology engineering part is one of the most work-intensive parts of semantic web research, and involves a lot of interaction with domain experts. Resulting ontologies, such as IFC, BOT, SSN, SAREF, and so forth are then highly valued, as they are community efforts from groups of domain experts defining their area of expertise. Now that patterns and association rules in building performance data are available in a graph, such data ideally also is evaluated by human domain experts, thereby endowing it with the necessary domain knowledge. Indeed, despite the ability to define objective knowledge (e.g. geolocations, element types, product data, etc.) because of the richness of ontologies, machines have considerable limitations when the data is highly dependent on context, subjective interpretation or is related to processes that are better performed by humans [28, 29, 30]. Such highly subjective cases require semantic contextualization, disambiguation, interpretation, similarity matching, etc. and are an essential aspect related to the richness of the semantic networks [30].

However, traditional methods of annotation by experts and semantic web technologies in general are based on the vision of a single correct truth, which does not fit the need of statistical validity and objectivity needed with data annotation. The "crowd truth" concept hereby aims to counteract subjectivity with the notion that interpretation gathered from a crowd will reduce subjectivity, provide more meaningful representations and reasonable interpretations [31]. In other words, subjective knowledge has no ground truth but relies on the dominant human opinion, which can be collected from the (expert) crowd [30].

[10] https://www.w3.org/community/lbd/
[11] https://w3id.org/bot#

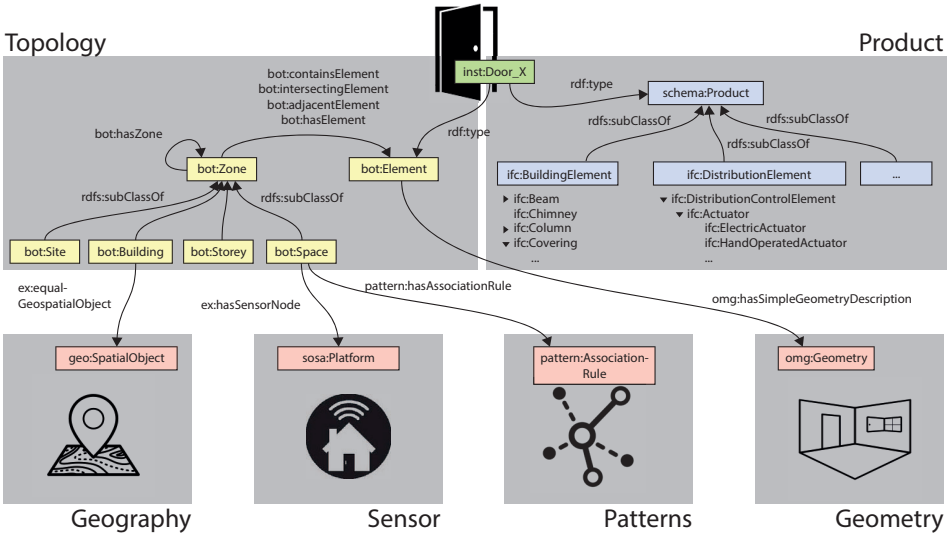

Figure 2: Conceptual overview of the modules and ontologies in a linked building data cloud, based on the work in the W3C LBD CG.

In the context of this research effort, it is essential to stress on the importance of the domain expertise. Interpretation of building performance patterns requires specific knowledge, so crowdsourcing is used in that context. The expert crowd in this case consists of professionals with high-level expertise in indoor environmental quality and building performance, who are highly familiar with the performance-oriented design process.

Howe [32] coined the term crowdsourcing and defined it as *"the act of a company or institution taking a function once performed by a designated agent (usually an employee) and outsourcing it to an undefined and generally large network of people in the form of an open call"*. According to several researchers, including Chiu et al. [33], the concept originates in research on open innovation and co-creation. Crowdsourcing techniques thereby allow to access and collect human intelligence and knowledge that are otherwise dispersed [34]. Surowiecki [35] states that the collective intelligence of the crowd will converge to a much more accurate solution than in the cases where experts contribute individually. According to the author, that is particularly valid when the contributors do not communicate with each other [35].

As a result, crowdsourcing has received major attention in the last decade in the areas of image recognition, product design and fabrication, rating systems, web development, etc. One of the most notable applications of such technologies is in design, including such based on AI techniques, where crowdsourcing combines human creativity with the machines' computational ability to explore various design proposals and solutions [36]. In the Semantic Web domain, crowdsourcing has been applied to collect high quality semantic annotations of data [28]. It has also been established as a way to obtain a sufficient number of human users for qualitative evaluation tasks [37]. Research on crowdsourcing in the context of the Semantic Web also indicates that crowdsourcing techniques are often used for ontology engineering and knowledge curation, knowledge validation, quality assurance of linked data, as well as crowd reviews and recommendations [38].

Research in the AEC domain has briefly touched upon the potential of crowdsourcing approaches in several different contexts. That includes the use of crowdsourcing techniques for extension of BIM-based construction material libraries through annotation of photos from site logs [39] and annotations

of construction workers based on video streams from building site [40]. The infrastructure domain has also realized some of the potential of crowdsourcing, as it has been used for co-creating and updating as-built BIM models, retrieving infrastructure operational and infrastructure condition information, co-creating sustainable and resilient infrastructure, as well as maintenance and rehabilitation [41].

From a technical perspective, [42] differentiate between various types of crowdsourcing platforms based on several criteria. The main difference stems from the diversity of the contributions and the ways in which they are aggregated. In terms of diversity of contributions, crowdsourcing platforms are defined as homogeneous (characteristically identical crowd contributions) and heterogeneous (contributions from the crowd differ significantly in nature and quality). In terms of aggregation, research contributions are divided in selective (value is extracted from individual contributions) and integrative (value is extracted from all contributions as a whole [42].

Considering that the context of this research requires high quality contributions from as many experts as possible, the methodological choice clearly indicates the need of a crowdsourcing platform relying on characteristically identical crowd contributions with value extracted from the entirety of all contributions. Blohm et al. [42] define this as "Information Pooling", which is a crowdsouricng technique aiming to aggregate distributed information and diverse opinions, assessments, predictions, etc. from contributors. The remainder of this article will, therefore, aim at implementing this form of crowdsourcing for the interpretation of building performance patterns by an expert crowd.

3. Methodology

In this article, we rely on a method proposed earlier in Petrova et al. [7, 9] for combining the results of KDD processes with semantic graphs of the building. This method is applied on a nearly zero energy building located near the city of Aarhus in Denmark. Using the open source SPMF data mining library, frequent repetitive patterns in the data (motifs) and association rules are discovered in the collected data for all spaces and represented using semantic data modelling techniques. The result is a performance-enriched semantic graph, which also includes building information, sensor placement, observed variables and sensor observations.

In addition, we devise a crowdsourcing web platform where discovered patterns can be semantically annotated by indoor environmental quality and building performance AEC experts. The aim is to retrieve interesting patterns identifying valuable hidden knowledge and filter out obvious dependencies. Patterns can be tagged (classified with semantic tags), so the system in the background can also classify them accordingly. The next two sections will document the (1) use case description, (2) implementation results for KDD and semantic data modelling for the use case, and (3) the implemented crowdsourcing platform.

4. Implementation

In this section, we document the overall concept implementation, thereby covering the use case description, applied KDD and semantic modelling approaches to retrieve the needed motifs and association rules, and the implementation of the crowdsourcing platform for their semantic annotation and classification.

4.1. Use case description

Home2020 ($132m^2$) is a detached house near the city of Aarhus, Denmark (Fig. 3), which was completed in 2017 and rated as nearly zero energy building (NZEB) according to the Danish energy labelling standard. The NZEB consists of a kitchen, a master bedroom, a living room, three additional rooms, two bathrooms, a utility room and a walk-in closet. The building occupants are a young working couple without children. The heat supply to the building is provided by district heating and distributed to a floor heating system. The hot water production and ventilation with heat recovery (85%) are supported by an air-to-water heat pump integrated in a compact unit. The ventilation system allows controlling the air supply in the living room and bedrooms individually and on-demand. The same applies to the extraction of air in the kitchen, bathrooms and the utility room. The air inlet is adjusted according to the levels of CO_2 and relative humidity in the rooms. Automatically controlled natural ventilation grids and skylights also allow to work towards optimal indoor environmental quality and thermal comfort while enhancing energy efficiency. The ventilation unit is running

with a minimum airflow when the house is unoccupied and when the indoor environmental conditions do not require a higher air supply. The ventilation system is automatically deactivated when the windows and doors are open. External solar shading devices are available in the living room and bedroom and can also be automatically controlled.

Figure 3: The Home2020 building.

A BMS is tracking several different performance parameters. That includes energy consumption is measured for the heating [MWh], ventilation system [kWh], control system [kWh], and kitchen appliances [kWh]. Other records also include outdoor air temperature [°C], return air temperature [°C], return air relative humidity [%], hot water temperature [°C], supply air temperature [°C], ventilation speed [steps]. Both hot and cold water consumption [m^3] are also monitored. In terms of indoor environmental quality, sensors monitor temperature [°C], CO_2 [ppm], and relative humidity [%]. The data is collected with a measurement interval of five minutes and the used dataset is from the period 01.12.2017 to 31.10.2018.

4.2. Knowledge discovery and semantic modelling of operational building data

This section presents the results from the motif discovery and association rule mining in the data from Home2020, as well as their semantic representation according to the methods described in the initial studies [7, 8, 9]. Additionally, the Home2020 use case building is also modelled using semantic data modelling techniques, and results in an RDF graph that is compliant with the ontologies and modelling recommendations set out by the W3C LBD CG. For reference, Listing 1 shows all namespaces and URIs (Unique Resource Identifiers) used in this effort.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    .
@prefix bot: <https://w3id.org/bot#> .
@prefix geo-ext: <http://eapetrova.com/voc/geoextension#> .
@prefix bmeta: <http://eapetrova.com/voc/buildingmetadata#>
    .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix ssn: <http://www.w3.org/ns/ssn/> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix om: <http://www.ontology-of-units-of-measure.org/
  resource/om-2/> .
@prefix ptn: <http://eapetrova.com/pattern/> .
@prefix list: <https://w3id.org/list#> .
@prefix inst: <https://home2020.dk/instances/> .
@prefix users: <https://home2020.dk/users/instances/> .
@prefix alltags: <https://home2020.dk/allTags/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix schema: <https://schema.org/> .
@prefix seas: <https://w3id.org/seas/> .
```

Listing 1: All namespaces used in the RDF graph.

First, the building itself has been modelled as an RDF graph according to the BOT ontology. This graph contains the description of building, building storeys and spaces. Also latitude, longitude, and altitude of the building are included using geospatial ontologies, as well as an OpenStreetMap (OSM) location[12]. The `ssn:hasProperty` predicate links each of the spaces to the sensor observations that are measured inside. Furthermore, the `bot:containsElement` containment relation relates the space to its contained sensor node.

In addition to the semantic data modelling, a lot of building performance patterns have been discovered from the operational building data using the SPMF open source data mining library for motif discovery and association rule mining. The collected measured building data is distributed over 335 log files in total, each of which is available as a CSV file containing the sensor data for one day. Data cleansing and data preparation were performed. The available data was analysed for erroneous data and outlier values, after which five iterations of multiple imputation were performed for the removal of missing values. An in-house software tool has been developed to be able to implement the complete KDD and semantic representation procedure. The procedure starts with parsing and loading all cleansed CSV data in memory, which, in the case of Home2020 comprises a total of 94434 measurements.

Symbolic Aggregate Approximation (SAX) is applied to all sensor observations for dimensionality reduction [43, 44]. SAX representations were generated on an hourly basis with the number of symbols equal to seven using the SPMF open-

[12] https://www.openstreetmap.org/

source data mining mining library[13]. As a result, all sequences of sensor data are transformed into SAX representations (strings of SAX symbols), each of which symbolizing one of the identified seven symbols (e.g. symbol '6' is equal to the interval [25.60784205790132,26.442700236915222] for the Temperature sequence). The complete sequence of data points is thus replaced by a symbolic representation similar to *32222223222222223333...* for each observed variable. Using the symbolic representations as an input, a matrix indicating the co-occurrence of the SAX symbols on a per month basis for all rooms and observed variables is also computed.

The identified co-occurrence matrices make it possible to further identify the frequent repetitive patterns (motifs) in the data (Longest Repeated Substrings (LRS)) with an implementation of the Suffix Tree algorithm [45]. The repeated substrings are the motifs or 'patterns', the dependencies between which form the association rules [46]. A manual data cleansing step is included at this point of the process to remove redundant data, i.e. overlapping patterns, patterns contained in each other, etc. The resulting motifs are then used to compute the co-occurrence matrices that show which patterns co-occur at any moment in time. As a result, the association rules can be derived, namely, rules that indicate the relations between co-occurring patterns in the different observed variables. The following Association Rule Mining (ARM) step is performed with an implementation of the CP-Growth algorithm and results in hundreds of association rules discovered in the observed variables in the different rules. Listing 2 shows some of the association rules that have been obtained as a result, including their measures of "interestingness": support and confidence. Finally, this output serves as the necessary input for the main contribution of this article: the semantic annotation and classification of association rules discovered in operational building data through crowdsouring techniques.

```
452 ==> 489 #SUP: 1 #CONF: 1.0
453 ==> 485 #SUP: 3 #CONF: 0.6
454 ==> 481 #SUP: 1 #CONF: 0.5
456 ==> 484 #SUP: 2 #CONF: 0.6666666666666666
457 ==> 488 #SUP: 1 #CONF: 1.0
```

Listing 2: Some of the association rules obtained for the living room in August.

[13]http://www.philippe-fournier-viger.com/spmf/

Each rule contains the IDs of the motifs that constitute the rule and the numerical value for support and confidence of the rule. The level of support hereby equals the number of co-occurrences that contains both the antecedent and consequent of the rule (the number of times the rule appears throughout the dataset). The confidence of a rule is an expression of how often that rule is found to be true. For example, if we consider rule 453 ==> 485 in the example set of results, motifs 453 and 485 co-occur 3 times (support = 3), with a confidence of 0.6. That means that only in three out of five times (only in 60% of all occurrences), pattern 453 co-occurred with pattern 485.

Figure 4 visualises that dependence for the same association rule 453 ==> 485. Patterns 453 and 485 represent two different SAX strings, namely 55544 (Humidity) and 5555544444 (Temperature). In other words, the rule indicates a relationship between the behaviour of the Humidity and Temperature observed variables. The symbols in the SAX strings hereby represent the precise intervals found earlier in the SAX computation step. For humidity, the SAX symbol '4' represents the interval [39.05,41.61] and '5' represents the interval [41.61,44.39] (percentage humidity). For temperature, the SAX symbol '4' represents the interval [24.73,25.35] and '5' represents the interval [25.35,26.03] (degree Celcius). In other words, whenever the indicated interval sequence in humidity occurs, there is a 60% chance that the corresponding interval sequence in temperature occurs as well.

All association rules and motifs are added to the semantic graph for the building using a built-for-purpose pattern ontology (`ptn:`). This ontology allows to represent the discovered association rules, including their `ptn:confidence`, `ptn:absoluteSupport`, and `ptn:relativeSupport`. The association rules are linked to individual sensor nodes using `ptn:hasAssociationRule` predicates. An indication of the resulting graph in RDF Turtle serialisation can be found in Listing 3.

```
inst:sensorNode_Kitchen
  rdf:type sosa:Platform ;
  sosa:hosts inst:Kitchen-CO2-Sensor, inst:Kitchen-
    Temperature-Sensor, inst:Kitchen-Humidity-Sensor ;
  ptn:hasAssociationRule inst:associationRule_1, inst:
    associationRule_2 .

inst:Kitchen-CO2
  rdf:type sosa:ObservableProperty .
```

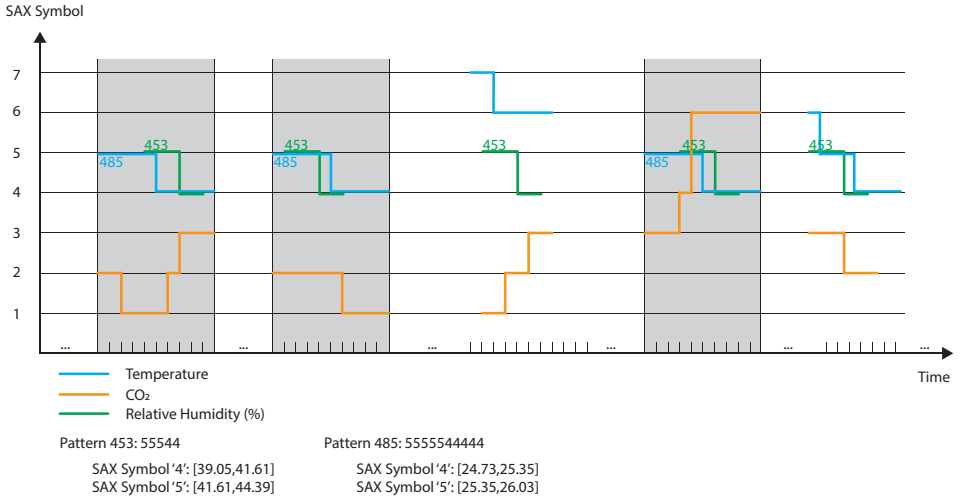

Figure 4: Diagram showing a number of association rules, leading to ARM support of 3 and confidence 0.6.

```
inst:Kitchen-CO2-Sensor
  rdf:type sosa:Sensor ;
  ssn:observes inst:Kitchen-CO2 .

inst:Kitchen-CO2-Sensor-obs1
  rdf:type sosa:Observation ;
  sosa:hasFeatureOfInterest inst:Kitchen ;
  sosa:hasResult [ a om:Measure ;
  om:hasNumericalValue "809.0"^^xsd:double ;
  om:hasUnit om:partsPerMillion ] ;
  sosa:madeBySensor inst:Kitchen-CO2-Sensor ;
  sosa:observedProperty inst:Kitchen-CO2 ;
  sosa:resultTime "01/12-2017 00:00:47"^^xsd:dateTime .

inst:associationRule_1
  rdf:type ptn:AssocationRule ;
  ptn:LHS (inst:Motif_45) ;
  ptn:RHS (inst:Motif_137) ;
  ptn:confidence "0.5"^^xsd:double ;
  ptn:absoluteSupport "1"^^xsd:double ;
  ptn:relativeSupport "0.5"^^xsd:double .

inst:motif_45
  rdf:type ptn:Motif ;
  ptn:SAXsequence "11122"^^xsd:string ;
  ptn:space inst:Kitchen ;
  ptn:month "8"^^xsd:string ;
  ptn:SAXsequenceFull (inst:SAXSymbol_91983cb8-4dd3-4544-
    a1fe-7b177e237bc0 inst:SAXSymbol_91983cb8-4dd3-4544-
    a1fe-7b177e237bc0 inst:SAXSymbol_91983cb8-4dd3-4544-
    a1fe-7b177e237bc0 inst:SAXSymbol_41fadfdb-6560-4e96-9
    a7f-bc405f453452 inst:SAXSymbol_41fadfdb-6560-4e96-9a7f
    -bc405f453452 );
  ptn:observedVariable "CO2"^^xsd:string .

inst:SAXSymbol_36ef82d8-57c9-4e0a-a0bc-c1c66404b02b
  rdf:type ptn:SAXSymbol ;
  ptn:symbol "5"^^xsd:int ;
  ptn:lowerBound "645.651281059915"^^xsd:double ;
  ptn:upperBound "700.959674546294"^^xsd:double .
```

Listing 3: RDF graph for the Home2020 building.

The identified association rules are very valuable for informing future design decision-making processes. Such insights can allow higher level performance analyses and can redefine the way decisions are taken in terms of, for example, spatial design, HVAC system design, considerations related to size of glazed areas in buildings, ventilation rates, prevention of overheating, optimal occupant comfort, etc.

The discovered motifs and association rules are at this point in the process embedded in and accessible from the RDF-based knowledge graph. Yet, considering the nature of the output from motif discovery and ARM, only the standard numerical expressions and measures are available, which do not convey any explicit semantics. To have an impact on decision-making, the discovered knowledge still has to be presented to and interpreted by a domain expert to identify the meaning, effects and implications of the discovered dependencies. The following sections use the results from the above summarised knowledge discovery process and indicate how input from domain experts can be retrieved and included in the knowledge graph. A distinction is hereby made between the semantic annotation of building performance patterns itself and the use of crowdsourcing techniques for both annotation and classification of the patterns according

to level of interestingness based not only on the numerical measures, but also domain expertise. Using those two steps together enables the transformation of the discovered knowledge into a decision support mechanism.

4.3. Crowdsourcing domain expertise

In this section, an overview is given of the proposed crowdsourcing platform and domain expert retrieval system. The dataset documented in the last section is used to demonstrate how the system works. A lot of contextual information is available about the building, including weather data, HVAC system data, occupant data, and so forth, in addition to the building topology, geospatial and operational data. An indication of what (a part of) the graph containing all these types of data looks like in GraphDB[14] is shown in Fig. 5. Adding these additional types of available data allows to present the discovered knowledge in the right context, which is a prerequisite for the provision of interpretation or causalities.

Furthermore, an overall system architecture diagram is displayed in Fig. 6, showing how all building data is made available with a linked data oriented interface (bottom right). Through a web-based crowdsourcing tool, which relies on a database of user profiles (experts), linksets and metadata are collected, which is the main purpose of the proposed tool.

The next sections indicate how input from domain experts can be retrieved and included in the knowledge graph to interpret the meaning of entities such as the rule described above. A two-step methodology is implemented, which first requires the semantic annotation of building performance patterns and then relies on crowdsourcing and the provided annotations to classify the building performance patterns. Both techniques are thereby employed together step-wise as part of the same crowdsourcing system.

4.3.1. Principles and ontologies

When experts are presented with an association rule, they can identify certain features and annotate the rule directly, as part of the semantic graph. Thus, original data, discovered performance patterns and rules, and interpretation by annotation are all stored in the same place, together with any additional contextual information or links to external information. In our crowdsourcing setup, a key aim is to have both semantic classifications and human-readable descriptions in the graph. This leads to the inclusion of more formal, rigorous and machine-oriented tags, *in addition to* the more ambiguous, informative, and human-oriented interpretations and descriptions that represent and convey the explanations of the occurring patterns and rules.

The expert annotations are gathered through the crowdsourcing platform and stored directly in the semantic building graph, thereby serving as a reference. Ideally, the system hereby relies on available and proven ontologies for the user annotations. The expert annotations hereby would typically lead to the addition of classifications, and/or to the addition of pattern clarifications and more descriptive comments. Whereas the former option is much more re-usable by a machine and very useful for information retrieval, the latter is more informative to a human user. That is due to the fact that descriptions include a more elaborate textual interpretation of the pattern or the rule. Such descriptions can, however, only be fully utilised by a human end user, whereas the machine would function optimally with explicit semantic tags.

When it comes to storing of the semantic classification tags and descriptions, a number of options are available. It is possible to use of the Review ontology[15], which provides classes as `Comment`, `Review`, `Feedback`. Essential in this case is that the ontology allows to link a `Review` and a "work" directly. That "work" is not formalised within this ontology and can represent any given or user-defined concept (in this case association rules). The `Review` is then the central concept of the ontology, and more details can be added, e.g. comments and feedback to the review. Agents or human users are thereafter defined using the FOAF ontology[16]. The targeted crowdsourcing effort (semantic classification tags and descriptions) can be implemented using the Review ontology, however, it does not allow representing predefined tags nor does it provide an option to build such a library of tags. Hence, even though it is possible to add reviews, comments and feedback, support for formally structured semantically meaningful tags is missing.

As an alternative, it is possible to use the

[14] http://graphdb.ontotext.com/

[15] http://vocab.org/review/

[16] http://xmlns.com/foaf/0.1/

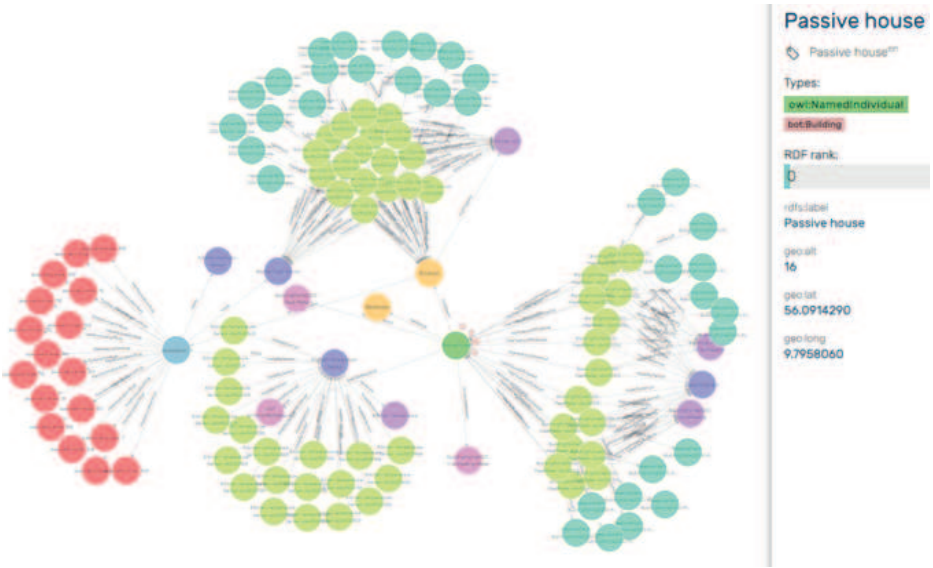

Figure 5: Contextualized graph for Home2020.

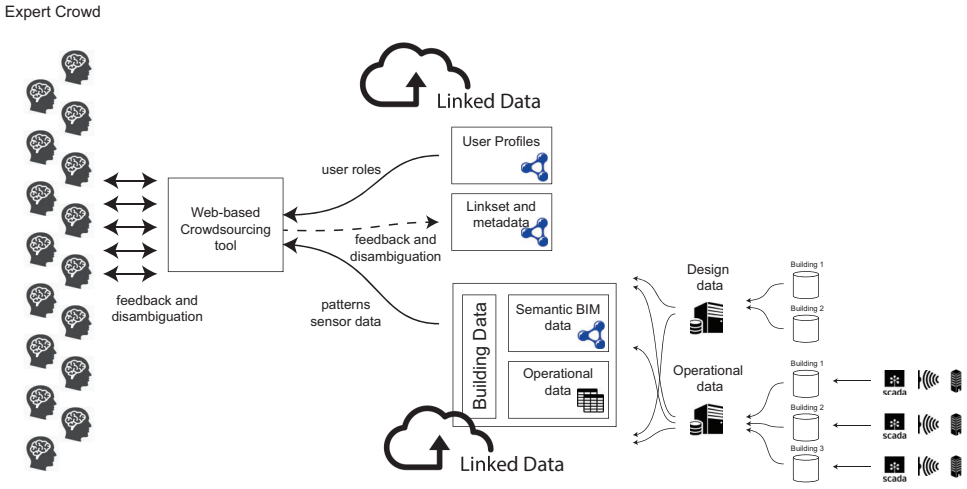

Figure 6: Overall system architecture diagram.

Review and Commenting definitions from the schema.org ontologies[17]. `Reviews` and `Comments` can in this case be linked to a `CreativeWork` directly. A `CreativeWork` is hereby defined as *"The most generic kind of creative work, including books, movies, photographs, software programs,*

[17] https://schema.org/Review

etc.”. Instead of using the FOAF ontology for modelling people, this ontology allows to use the `schema:Person` class. It also provides the option to store votes (e.g. `schema:upvoteCount`), which is of particular value to the envisioned crowdsourcing system. This ontology also allows to combine `Reviews`, `Comments`, and `CreativeWorks` in various ways. Finally, it is also possible to add metadata for each of these three concepts (agent, about, dateCreated, text, etc.).

Applying the schema.org ontology to the proposed semantic tagging and annotation system leads to a data model as displayed in Fig. 7. We suggest to use both `foaf:Agent` and `schema:Person` nodes for the representation of people data. This data constitutes the user profile database shown in the upper right area of Fig. 6.

So far, the proposed approach only makes it possible to add reviews with freely written text (human-readable descriptions). This approach is useful, but it does not provide the necessary semantically definitive tags or classifications, which are needed to be able to retrieve information in a machine-readable form. Therefore, in a next step, the proposed system is further extended with the possibility to add semantically defined tags (classification).

4.3.2. Semantic annotation tags

First and foremost, the implementation of a semantic tagging system requires the definition of a logical structure in terms of tagging classes, categories, labels and argumentation for the choice of those. In this case, five main categories can be used to group or classify tags that reflect the most usual causes of any regularity or irregularity appearing in operational building data. These are related to dynamic parameters that have a direct effect on building performance. These constitute the main classification tags: (1) external conditions, (2) occupant behaviour, (3) system performance, (4) design and (5) construction. During classification and tagging of the association rules, domain experts are able to associate comments to any of those categories and, if needed, define new ones.

Storage of the targeted tags does not rely on any of the formal ontologies discussed in the previous section. Apart from the five main tag categories discussed above, all other tags should ideally come from the crowd accessing the system. An `AllTags` ontology is thereby built in support of the entire crowdsourcing setup. We hereby recommend adopting a data dictionary approach, in which content can be added to a global dictionary, depending on acceptance of proposals coming from the crowd by a curator group.

Under each of the five main classification tags, a number of standard tags can thus be made available through the `AllTags` ontology. Any of these tags can be selected by the domain expert for annotation of an association rule. Furthermore, the system allows to add new, previously undefined tags, as deemed necessary by the domain expert. Over time, the number of default available tags can be revised by the curator group in charge of the `AllTags` ontology or tags dictionary, in order to better respond to the tagging behaviour and requirements.

The tags need to be collected and stored, so this work approaches this by storing all tags into a separate graph, to which additional tags can be added as preferred. Ideally, a user does not need to devise new tags continuously, but instead can rely on the tags available in this `AllTags` ontology. As a result, a number of tags are available under each category (see `subClassOf` tree structure in Fig. 8). These tags can be presented to and ticked by domain experts for the semantic annotation of their reviews of building performance patterns.

4.3.3. Platform-User interaction

The previous sections defined the data model that can be used for semantic annotations and classification (tagging) of building performance patterns by domain experts. Naturally, to be fully useful, this data model needs to be embedded in a fully implemented web-based application that presents domain experts with patterns and rules and allows them to provide their input. Even though the documentation of that application and its user interface are out of scope of this article, an interaction diagram can be provided that indicates how feedback and comments are retrieved (Fig. 9).

As shown in the interaction diagram in Fig. 9, ARM nodes identified with URIs are retrieved from the graph. The relevant contextual information is also included at that stage (Steps 1-3). If reviews are already available for the selected instance, they are presented to the user as well. This gives the domain expert the opportunity to upvote already available reviews, depending on whether or not the reviews are considered to be correct, or indicative of a particular level of “interestingness” (Step 4a). At any time during that process, the expert user is able to assign a new review. Metadata is then attached

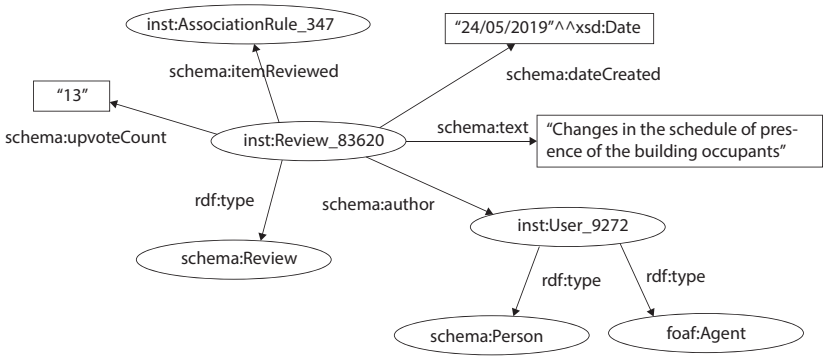

Figure 7: RDF graph for annotating an association rule with reviews, descriptions, metadata, and votes.

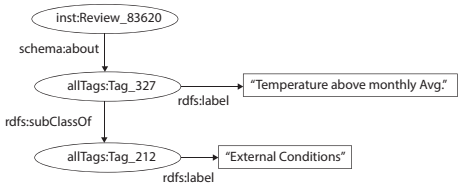

Figure 8: Adding Tags to the Tag database.

to each new review (user metadata, date, profile) (Step 4b) together with a description and a semantic tag (from the repository of tags) (Steps 5a, 5b). This work proposes to store all reviews and comments in a separate graph, and link those to the corresponding association rule URIs and the user URIs, as indicated in the previously defined data model. In terms of the system architecture, all reviews and comments are stored in the "Linkset and metadata" database (middle right in Fig. 6), including the URIs of association rules in the LBD graph (bottom right in Fig. 6) and the URIs of people in the user profile database (upper right in Fig. 6).

For each Review tag or description added to an association rule by a domain expert with user details available after login, a new "Review" node is added, including the associated user profile, a date, and a human-readable description, as can be seen in Fig. 9. Thus, motif and ARM nodes in the graph are retrieved together with additional metadata, i.e. classification tags, user metadata, user profile, and so forth.

4.3.4. The effect of the crowd

The semantic tagging mechanism outlined above is only as good as the input tags and human descriptions. For an end user, or for a system, it is still very difficult to find out which patterns are of higher or lesser interest. Alternatively, a semantic system might be built that focuses less on the semantic annotations and more on the classification of rules according to the numerical values of the 'interestingness' measures. At this point, without any semantic annotation, all the association rules in the data are rather similar, with differences mainly in the support and confidence. Instead of adding specific semantic annotations, as outlined above, a useful alternative may be to to let domain experts take a completely unsupervised approach, and browse association rules without using any predefined method. Due to the nature of human expertise, it might be sufficient to indicate pattern co-occurrences (or association rules) and devise which the interesting ones are based on browsing behaviour (no semantic tags or descriptions).

Therefore, the system outlined in the previous sections is extended with a crowdsourcing mechanism focusing on interest among domain experts. Although this could be seen as a separate crowdsourcing tool, it is added to the above outlined system for semantic annotation. An overall diagram of the data that could be produced by such a joint system, is provided in Fig. 10. The top of this diagram shows the semantic annotation mechanism of the previous section, for which an interaction diagram was presented in Fig. 9. The lower part of this diagram shows the mechanism for upvoting of Reviews.

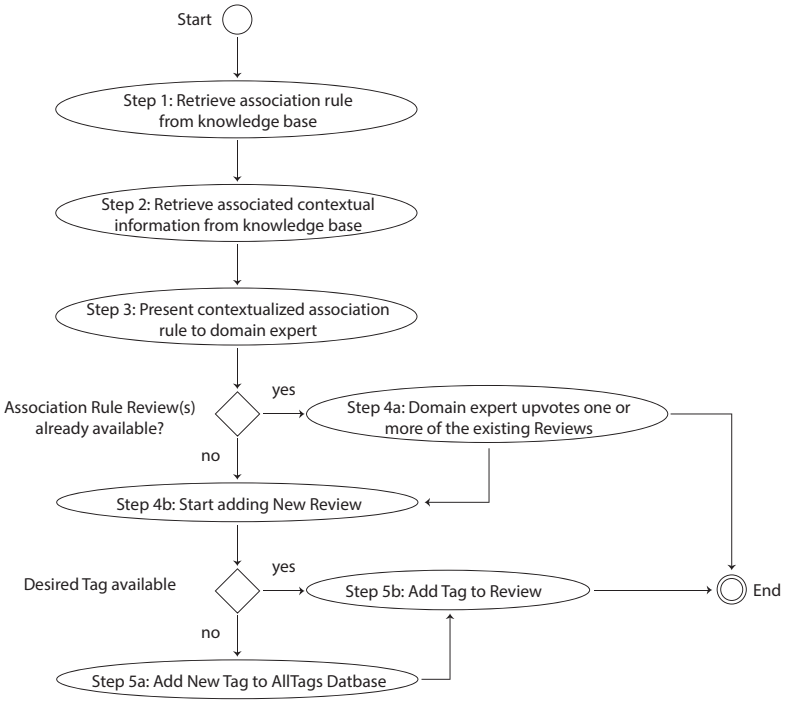

Figure 9: Interaction diagram for semantic annotations.

The data model in Fig. 10 shows the proposed implementation of an upvoting mechanism for association rules as well (inst:AssociationRule_347 schema:upvoteCount "12").

The direct upvoting of semantic rules, without the addition of semantic tags and descriptions, requires little effort and thus has the potential for a lot of data generation, in addition to the more work-intensive semantic tagging and disambiguating description procedure discussed earlier. When a user logs in, and activates their user profile, they can browse the available association rules. Based on expertise, the expert is able to identify the most interesting patterns and indicates this accordingly for the association rule. This direct tagging mechanism provides an indication of popularity and interestingness, which may be a valuable addition to the main rigorous crowdsourcing-based disambiguation mechanism.

5. Results

5.1. The crowdsourcing system

The presented linked data based system allows domain experts to contribute in interpreting patterns and association rules using a crowdsourcing approach. Because of its setup with a number of feedback options, as indicated in Section 4, the following section summarises the three main expert crowd contributions:

1. Input:
   Domain experts (User 1 and User 2 in Fig. 10) provide their input about new association rules or refine and update already existing crowd contributions. The users choose which rules to engage with without predefined suggestions or other constraints. Once provided, the expert input is stored in the semantic graph.

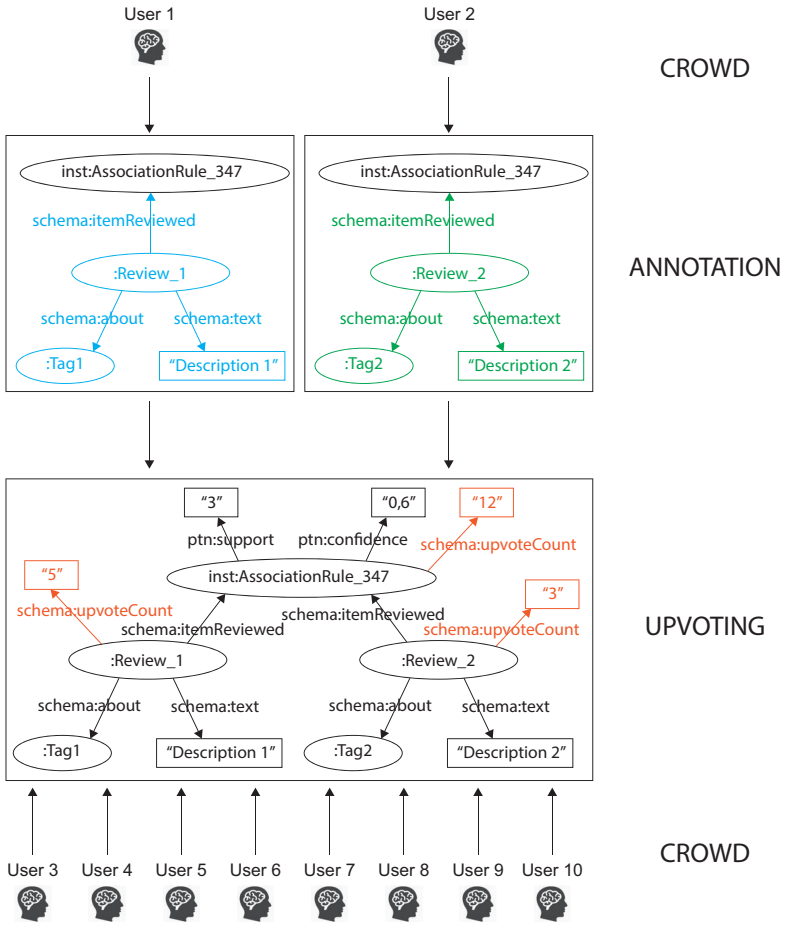

Figure 10: Interaction diagram for upvoting.

2. Review:
Other members of the expert crowd (Users 3-10 in Fig. 10) also engage with the system by annotating and tagging new rules, or interacting with existing reviews, thereby upvoting or refining previous annotations. That input is also stored and analysed against the existing semantic graph. The system provides real-time feedback about any inconsistencies during the update.

3. Upvote:
The users vote on triples suggested by other users. An annotations is upvoted in case the expert's belief is similar to an existing annotation from another expert.

5.2. Collected data

The system has only been tested in an alpha state, meaning that its main functionalities and approach have been tested for viability and consistency. A full implementation with full beta-testing is out of scope for this work, and is targeted at in a next phase of the research. For the alpha testing phase, the building data and patterns for the

Home2020 case have been extended with example expert input data (Listing 4) for the same association rule that has been used throughout the article. Upvote counts are taken from the example displayed in Fig. 7 and 10. This Listing includes all the kinds of input that can be provided by domain experts, as outlined in the previous section, i.e. input, reviews, and upvote.

```
inst:AssociationRule_347
  rdf:type ptn:AssociationRule ;
  ptn:LHS (inst:Motif_453) ;
  ptn:RHS (inst:Motif_485) ;
  ptn:confidence "0.6"^^xsd:double ;
  ptn:absoluteSupport "3"^^xsd:double ;
  ptn:relativeSupport "0.6"^^xsd:double .

inst:Review_83620
  a schema:Review ;
  schema:itemReviewed inst:AssociationRule_347 ;
  schema:dateCreated "24/05/2019"^^xsd:date ;
  schema:text "Changes in the schedule of presence of the
    building occupants"^^xsd:string ;
  schema:about alltags:Tag_6251 ;
  schema:author users:User_9272 ;
  schema:upvoteCount "5" .

inst:Review_32486
  a schema:Review ;
  schema:itemReviewed inst:AssociationRule_347 ;
  schema:dateCreated "12/04/2019"^^xsd:date ;
  schema:text "Ventilation system malfunctioned"^^xsd:
    string ;
  schema:about alltags:Tag_324 ;
  schema:author users:User_316 ;
  schema:upvoteCount "3" .

inst:AssociationRule_347
  schema:upvoteCount "13" .

alltags:Tag_324
  a owl:Class ;
  rdfs:subClassOf alltags:Tag_3 ;
  rdfs:label "System malfunction" .

alltags:Tag_6251
  a owl:Class ;
  rdfs:subClassOf alltags:Tag_2 ;
  rdfs:label "Schedule change" .

alltags:Tag_2
  a owl:Class ;
  rdfs:label "Occupant behaviour"@en .

alltags:Tag_3
  a owl:Class ;
  rdfs:label "System performance"@en .
```

Listing 4: RDF graph with reviews and votes.

Note that user data as well as tags are maintained in separate RDF graphs with distinct URIs. Furthermore, using SPARQL queries, the necessary information can be retrieved to obtain the relevant information for visualisation in the crowdsourcing tool. An example query is provided in Listing 5, showing how the `schema:upvoteCount` can be retrieved for all association rules that have been commented on. Of course, many more diverse queries are possible, also for updating the system with new data (update of upvoteCount, adding a review, adding a tag in the AllTags database, etc.).

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX inst: <https://home2020.dk/instances#>
PREFIX ptn: <http://eapetrova.com/pattern/>
PREFIX schema: <http://schema.org/>
select ?ar ?t where {
  ?ar a ptn:AssociationRule .
    ?ar schema:upvoteCount ?t .
    ?rev schema:itemReviewed ?ar
}
```

Listing 5: SPARQL query allowing to retrieve the number of upvotes for a specific set of association rules.

6. Conclusion

6.1. Summary

Recent technologies have had a strong impact on decision support in the AEC industry. In addition to BIM innovations, semantic data modelling and machine learning techniques bring even stronger shifts and present more opportunities for design decision support. This article first briefly outlines how both sets of technologies can be combined to enable both the semantic representation of buildings (topology, occupants, geography, etc.), and retrieve motifs and rules discovered in operational building data. All this information can be combined into a single knowledge graph that is accessible for information retrieval.

A key challenge in the above, which has been outlined in this article, is the lack of semantics in the retrieved building performance patterns. Such knowledge is entirely statistical in nature, and their meaning and usefulness are unclear, which makes them less useful for decision support and less amenable for information retrieval. The only way to make discovered performance patterns useful to a decision-making process, is to have them explained by a domain expert. In order to obtain the highly necessary domain knowledge in a machine-readable form, this article looks into the use of crowdsourcing techniques for enriching discovered building performance patterns with interpretation from domain experts. This allows to build distributed knowledge graphs of building data, enriched with building performance patterns and interpreted by indoor environmental quality and energy performance experts. A proof of concept implementation is presented.

The proof of concept shows three main potential ways in which crowdsourcing techniques may be

used to endow building performance patterns with domain knowledge and interpretation. First of all, *input* may be targeted, in which case domain experts annotate data with reviews that link association rules to machine-oriented semantic tags and/or human-oriented descriptions. Second, *reviews* may be targeted, which are confirmations and additions by domain experts who upvote existing reviews or provide new reviews. Third, *upvotes* may be targeted, not only for the reviews, but also for the association rules themselves. Such upvotes provide little meaning (as opposed to the input and review options for feedback), but rather give a measure of interestingness.

6.2. Evaluation

The proposed crowdsourcing approach for interpretation and annotation of building performance patterns can be useful, as it allows experts to engage directly with the existing hierarchy of classes. The users do not have to be familiar with the existing semantic graph to provide new input. Furthermore, Semantic Web technologies and reasoning mechanisms can be valuable for analysing user input, assure quality and ensure that there are no contradictions between different annotations. Such a system can also serve as an educational mechanism for the domain-specific crowd.

Another important aspect is the accumulation of semantic annotations and tags over time. The semantic annotations and tags have to accumulate to a point in time where they become statistically significant and useful. To assure usefulness, the system and its functionalities have to be tested in a rel life experiment with domain experts.

6.3. Future work

The initial evaluations of the proposed system show that crowdsourcing techniques hold significant potential for interpretation and semantic annotation of knowledge discovered in operational building data. Such an approach can help with removing the unexplored space between traditional machine learning approaches for knowledge discovery and semantic data modelling for knowledge representation. However, some challenges need to be addressed in future work.

First of all, even though knowledge discovered in building performance data can be interpreted and retrieved, that by itself does not provide immediate decision support. To be of use, such a system has to be integrated in a specific design decision support system that targets the design professional directly and is able to impact their workflow and results in a positive way. Second, even though the crowd consists of domain experts, it has to be assumed that the quality of the contributions may vary. That requires the consideration of an additional quality assurance mechanism, which could be either an additional round of expert reviews or a rule-based system. Finally, the actual usefulness of the crowd annotations has to be evaluated. Not all patterns are equally valuable for design decision support, so an additional supporting mechanism that is able to filter out the valuable novel knowledge needs to be considered.

References

[1] C. M. Eastman, P. Teicholz, R. Sacks, K. Liston, BIM handbook: a guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators, John Wiley & Sons, Hoboken, NJ, USA, 2008.

[2] R. Sacks, C. M. Eastman, G. Lee, P. Teicholz, BIM handbook: a guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators, 3rd Edition, John Wiley & Sons, Hoboken, NJ, USA, 2018.

[3] A. Borrmann, M. König, C. Koch, J. Beetz, Building Information Modeling: Technology Foundations and Industry Practice, 1st Edition, Springer, 2018. doi:10.1007/978-3-319-92862-3.

[4] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American 284 (5) (2001) 34–43.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine 17 (3) (1996) 3754.

[6] C. Bishop, Pattern Recognition and Machine Learning, Springer, Hoboken, NJ USA, 2006, ISBN: 978-0-470-54137-1.

[7] E. Petrova, P. Pauwels, K. Svidt, R. Jensen, Towards data-driven holistic sustainable design: A decision support framework relying on knowledge discovery in real-time building performance data and disparate project data repositories, Architectural Engineering and Design Management (2018) 1–23.

[8] E. Petrova, P. Pauwels, K. Svidt, R. Jensen, From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches, in: Proceedings of the 11th European Conference on Product and Process Modelling (ECPPM), 2018, pp. 391–398.

[9] E. Petrova, P. Pauwels, K. Svidt, R. Jensen, In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data, in: Proceedings of the CIB W78 Conference, 2018, pp. 19–26.

[10] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, 2011.

[11] J. Han, M. Kamber, J. Pei, Data mining concepts and techniques, Morgan Kaufmann, Waltham, US, 2012.

[12] A. Lausch, A. Schmidt, L. Tischendorf, Data mining and linked open data – new perspectives for data analysis in environmental research, Ecological Modelling 295 (2015) 5–17.

[13] S. Shekhar, P. Zhang, Y. Huang, Spatial data mining, in: Data Mining and Knowledge Discovery Handbook 837-854., Springer, 2010.

[14] Z. Wang, R. Srinivasan, A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models, Renewable and Sustainable Energy Reviews 75 (2017) 796–808.

[15] S. D'Oca, T. Hong, A data-mining approach to discover patterns of window opening and closing behavior in offices, Building and Environment 82 (2014) 726–739.

[16] Z. Cheng, Q. Zhao, F. Wang, Z. Chen, Y. Jiang, Y. Li, Case studies of fault diagnosis and energy saving in buildings using data mining techniques, in: Proceedings of IEEE international conference on automation science and engineering, IEEE, Fort Worth, TX, 2016, pp. 645–651.

[17] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy and Buildings 75 (2014) 109–118.

[18] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Automation in Construction 49 (2015) 1–17.

[19] K. McGlinn, B. Yuce, H. Wicaksono, S. Howell, Y. Rezgui, Usability evaluation of a web-based tool for supporting holistic building energy management, Automation in Construction 84 (2017) 154–165.

[20] S. Yarmohammadi, R. Pourabolghasem, D. Castro-Lacouture, Mining implicit 3d modeling patterns from unstructured temporal bim log text data, Automation in Construction 81 (2017) 17–24.

[21] C. Bizer, T. Heath, T. Berners-Lee, Linked data - the story so far, International Journal on Semantic Web and Information Systems 5 (3) (2009) 1–22.

[22] P. Pauwels, S. Zhang, Y.-C. Lee, Semantic web technologies in aec industry: a literature review, Automation in Construction 73 (2017) 145–165. doi:10.1016/j.autcon.2016.10.003.

[23] P. Pauwels, W. Terkaj, EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology, Automation in Construction 63 (2016) 100–133, DOI: 10.1016/j.autcon.2015.12.003.

[24] M. H. Rasmussen, P. Pauwels, C. A. Hviid, J. Karlshøj, Proposing a central AEC ontology that allows for domain specific extensions, in: F. Bosché, I. Brilakis, R. Sacks (Eds.), Proceedings of the Joint Conference on Computing in Construction, Vol. 1, Heriot-Watt University, Heraklion, Crete, Greece, 2017. doi:10.24928/jc3-2017/0153.

[25] G. F. Schneider, Towards aligning domain ontologies with the Building Topology Ontology, in: 5th Linked Data in Architecture and Construction Workshop, University of Burgundy, Dijon, France, 2017. doi:10.13140/RG.2.2.21802.52169.

[26] G. F. Schneider, M. H. Rasmussen, P. Bonsma, J. Oraskari, P. Pauwels, Linked Building Data for Modular Building Information Modelling of a Smart Home, in: eWork and eBusiness in Architecture, Engineering and Construction (ECPPM 2018), CRC Press, Copenhagen, Denmark, 2018, pp. 407–414.

[27] K. McGlinn, A. Wagner, P. Bonsma, L. McNerney, D. O'Sullivan, Interlinking geospatial and building geometry with existing and developing standards on the web, Automation in ConstructionIn press.

[28] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, J. Lehmann, Crowdsourcing linked data quality assessment, in: H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. Parreira, L. Aroyo, N. Noy, C. Welty, K. Janowicz (Eds.), The Semantic Web – ISWC 2013, Vol. 8219 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2013, pp. 260–276.

[29] M. Acosta, Crowdsourcing linked data management, in: A. Bernstein, J. Leimeister, N. Noy, C. Sarasua, E. Simperl (Eds.), Crowdsourcing and the Semantic Web, Vol. 4 of Dagstuhl Reports, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, 2014, p. 29.

[30] H. Xin, R. Meng, L. Chen, Subjective knowledge base construction powered by crowdsourcing and knowledge base, in: Proceedings of the SIGMOD'18: 2018 International Conference on Management of Data, 2018, pp. 1349–1361, houston, TX, USA. doi:https://doi.org/10.1145/3183713.3183732.

[31] L. Aroyo, Semantic interpretation and crowd truth, in: A. Bernstein, J. Leimeister, N. Noy, C. Sarasua, E. Simperl (Eds.), Crowdsourcing and the Semantic Web, Vol. 4 of Dagstuhl Reports, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, 2014, p. 31.

[32] J. Howe, The rise of crowdsourcing, Wired Magazine 14, available at: https://www.wired.com/2006/06/crowds/ (accessed 27 May 2019).

[33] C. Chiu, T. Liang, E. Turban, What can crowdsourcing do for decision support?, Decision Support Systems 65 (2014) 40–49.

[34] E. Schenk, C. Guittard, Towards a characterization of crowdsourcing practices, Journal of Innovation Economics 7 (2011) 93–107.

[35] J. Surowiecki, The Wisdom of Crowds, Random House, New York, US, 2005.

[36] W. Xiang, L. Sun, W. You, C. Yang, Crowdsourcing intelligent design, Frontiers of Information Technology & Electronic Engineering 19 (1) (2018) 126–138.

[37] H. Sack, Crowdsourcing for evaluation and semantic annotation, in: A. Bernstein, J. Leimeister, N. Noy, C. Sarasua, E. Simperl (Eds.), Crowdsourcing and the Semantic Web, Vol. 4 of Dagstuhl Reports, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, 2014, pp. 43–44.

[38] C. Sarasua, E. Simperl., N. Noy, A. Bernstein, J. Leimeister, Crowdsourcing and the semantic web: A research manifesto, Human Computation 2 (1) (2015) 3–17. doi:10.15346/hc.v2i1.2.

[39] K. Han, M. Golparvar-Fard, Crowdsourcing BIM-guided collection of construction material library from site photologs, Visualization in Engineering 5 (14). doi:10.1186/s40327-017-0052-3.

[40] K. Liu, M. Golparvar-Fard, Crowdsourcing construction activity analysis from jobsite video streams, Journal of Construction Engineering and Management 141 (11). doi:10.1061/(ASCE)CO.1943-7862.0001010.

[41] S. Consoli, R. Reforgiato, An urban fault reporting and management platform for smart cities, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 535–540.

[42] I. Blohm, S. Zogaj, U. Bretschneider, J. Leimeister, How to manage crowdsourcing platforms effectively?, California Management Review 60 (2) (2018) 122–149.

[43] P. Patel, E. Keogh, J. Lin, S. Lonardi, Mining motifs in massive time series databases, in: 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 2002, pp. 370–377. doi:10.1109/ICDM.2002.1183925.

[44] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Mining and Knowledge Discovery 15 (2007) 107–144.

[45] P. Weiner, Linear pattern matching algorithms, in: 14th Annual Symposium on Switching and Automata Theory (SWAT 1973), 1973, pp. 1–11. doi:10.1109/SWAT.1973.13.

[46] T.-C. Fu, A review on time series data mining, Engineering Applications of Artificial Intelligence 24 (2011) 164–181. doi:10.1016/j.engappai.2010.09.007.

Appendix F. Paper VI

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2019,in press). Semantic data mining and linked data for a recommender system in the AEC industry. Accepted for publication in *Proceedings of the 2019 European Conference on Computing in Construction*, 10-12 July, Chania, Crete, Greece.

2019 EC³
2019 European Conference on Computing in Construction
Chania, Crete, Greece
July 10-12, 2019

SEMANTIC DATA MINING AND LINKED DATA FOR A RECOMMENDER SYSTEM IN THE AEC INDUSTRY

Ekaterina Petrova[1,*], Pieter Pauwels[2], Kjeld Svidt[1] and Rasmus Lund Jensen[1]

[1]Department of Civil Engineering, Aalborg University, Aalborg, Denmark
[2]Department of Architecture and Urban Planning, Ghent University, Ghent, Belgium
ep@civil.aau.dk*, pipauwel.pauwels@ugent.be, ks@civil.aau.dk, rlj@civil.aau.dk

Abstract

Even though it can provide design teams with valuable performance insights and enhance their decision-making processes, monitored building data is rarely reused in an effective feedback loop from operation to design. Data mining allows users to obtain such novel insights from the large datasets generated throughout the building life cycle. Furthermore, semantic web and linked data technologies allow to formally represent the built environment and retrieve knowledge in response to domain-specific requirements. Both approaches have independently established themselves as powerful aids in decision-making. Combining both can enrich data mining processes with domain knowledge and facilitate knowledge discovery, representation and reuse. In this article, we look into the available data mining techniques and investigate to what extent they can be fused with semantic web technologies, to provide recommendations to the end user in a performance-oriented design process. We demonstrate an initial implementation of a linked data-based system for generation of recommendations.

Introduction

Building data: BIM and semantic web technologies in a sensor world

Recent years have presented significant research efforts accentuating the environmental contribution from the built environment and methods for its mitigation. That has amended design and engineering practice and has made it strive towards implementing sustainability principles as fundamental and not merely complementary. Simultaneously, the rapid technological developments have allowed powerful computational techniques to emerge in support of architectural design and engineering. They allow to represent buildings semantically (El-Diraby 2013, Pauwels et al. 2017) and discover implicit knowledge about their performance through pattern recognition and knowledge discovery techniques (Fayyad et al. 1996). With regards to data representation in Architecture, Engineering and Construction (AEC), Building Information Modelling (BIM) allows the creation of semantically rich building models (Borrmann et al. 2018, Sacks et al. 2018).

Recently, semantic web technologies (Berners-Lee et al. 2001) have received major attention in the attempt to break open the isolated silos of information and connect the semantically rich building data with other meaningful data about the building, its occupants, environment, etc. These further reaching semantic models are the building blocks of Linked Building Data (LBD) and provide a decentralized source of information (Pauwels 2014). On the other hand, Building Monitoring/Automation Systems (BMS/BAS) play an essential role in building operation, by allowing the collection of operational data through a myriad of sensors and devices (Fan et al. 2015, Xiao & Fan 2014). Advanced analytical methods are hereby of high value, as they help uncover hidden knowledge in the data generated during operation, and highlight its potential to the future of building design and performance improvement (Molina-Solana et al. 2017, Miller et al. 2018).

Despite the availability of knowledge bases, many of the decisions taken during the design process are based on 'rules of thumb' and previous experience (Heylighen et al. 2007), and not on data and evidence contained in building performance, BIM models or LBD knowledge graphs. If such data were used more efficiently, significant potential would be uncovered in reaching performance targets currently associated with gaps between design and actual performance (ODonnell et al. 2013, Corry et al. 2015, de Wilde 2014). Precisely this is the target of this research effort: bringing knowledge from previous projects into future design environments to achieve both a sustainable end product and a holistic sustainable design process. Previous works also investigated how Knowledge Discovery in Databases (KDD) (Fayyad et al. 1996) can be used to retrieve patterns and association rules from available building data (Petrova et al. 2018*a*,*b*,*c*). These works also showed how it is possible to build a knowledge graph that includes (1) semantically rich building data (topology, product data, properties), (2) 2D and/or 3D geometry, (3) sensor data, and (4) motifs and association rules obtained from the sensor data. The resulting graph provides a valuable resource for evidence-based design recommen-

dations. Therefore, the objective of the current article is to investigate the potential of linked (open) data-based recommendation retrieval in the design environment, including patterns obtained through mining of sensor data, thereby utilizing the available and ever-growing knowledge bases to achieve an evidence-based design process.

Linked data-based recommender systems for improving sustainable design decision-making

In this work, we look into the possibility of building a system that relies on knowledge graphs to make recommendations towards the design team. Considered here is evidence-based feedback in response to design requirements, yet the recommender system is conceived as user-centered and can provide any feedback requested by querying the available knowledge base(s). Generally, recommender systems can be subdivided in content-based and graph-based (Musto et al. 2017). A content-based system hereby provides recommendations based on direct similarity. A graph-based one directly links user nodes to specifically user-tailored recommendations, to improve search and content retrieval.

Several research efforts investigate recommender systems based on linked data (graph-based) and the wealth of data provided by the Linked Open Data (LOD) cloud[1](Oliveira et al. 2017, Musto et al. 2017). Research in the area of LOD-based recommendations takes its roots in the field of ontology-based recommender systems, introduced by Middleton et al. (2004). When linked data and ontologies are used for the disambiguation of content, recommendation systems become semantics-aware (de Gemmis et al. 2015, Boratto et al. 2017). Early recommender systems typically combine linked data with closed systems, thereby aiming to improve recommendations with more structured and semantically richer user data (Heitmann & Hayes 2010). The use of linked data for user-centered recommendations was introduced by Passant (2010), who proposed a music recommender system based on semantic similarity calculations involving DBpedia[2] properties. This research relies on a set of measures to compute the semantic distance in linked data, thus exploiting the abundance of links among the resources. Most recent works (Oliveira et al. 2017, Boratto et al. 2017) typically follow the software architecture displayed in Fig. 1 where user profiling is on focus.

Recommender systems are usually associated with user profiling and suggestion generation based on previous interactions, social relations, likes, etc. In other words, recommendations aim for matching the user's demands (profile) with the highest possible level of similarity, while still diversifying the recommendations and not limiting to the same content. In the case of building design, the similarity matching aspect should also be the starting point, but it should be equally balanced with diversification driven by design and performance requirements. For example, if a user profile indicates high interest in residential nearly zero-energy buildings (NZEB), the recommender system should also be able to suggest different NZEB building types or other residential building types, etc. (diversification). Of course, the richer the original dataset, the easier it is to obtain and make alternative recommendations.

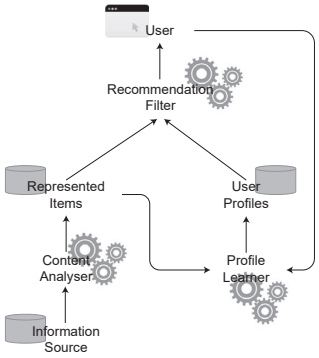

Figure 1: Semantics-aware content-based recommender system (based on Boratto et al. (2017))

Recommendation engines are not unknown to the AEC industry. However, these usually conceived to suggest predefined objects hosted in a database when a certain level of similarity with the current design is achieved (content-based). As a result, one of the fundamental goals of this research endeavor is to investigate the level of feasibility for application of linked data-based recommendations utilizing dynamic knowledge bases in changing context. In the same example, the dynamic knowledge bases are new buildings projects, which may include continuous incoming streams of sensor data and new LBD graphs. The changing context then refers to continuously updating user profiles.

To achieve the above-stated objectives, this paper starts with a state of the art review in the areas of KDD, semantic web technologies and data stream processing. The article then continues with the approach we use to achieve the objectives of the current study. We then outline the necessary steps towards a linked data-based recommender system for improvement of decision-making in sustainable building design and perform initial tests. Finally, the paper discusses the results, presents the main conclusions and outlines future work.

[1]https://lod-cloud.net/
[2]https://wiki.dbpedia.org/

State of the Art

Knowledge Discovery in Databases (KDD) according to data type and purpose

Fayyad et al. (1996) define KDD as the overall process, in which knowledge is the end product of data-driven discovery. They outlinee five main steps in that process, namely selection, pre-processing, transformation, data mining and interpretation/ evaluation of the results. In that context, Hand et al. (2011) define data mining as *"the analysis of large observational datasets to find unsuspected relationships and summarise the data in novel ways so that data owners can fully understand and make use of the data"*. Fayyad et al. (1996) also summarise six main data mining categories, i.e., classification, clustering, association rule mining, regression, summarization and anomaly detection. Han et al. (2012) divide these into two main categories: predictive (supervised) and descriptive (unsupervised). Descriptive analystics use data aggregation and mining to provide insight into the past and make it interpretable by humans. Predictive analytics use statistical models and forecasting approaches to understand the future and provide actionable insights. With regards to the input data source, Lausch et al. (2015) distinguishes predominantly between (numerical and categorical) data, text, web, media, time series and spatial data mining.

Knowledge discovery in Architecture, Engineering and Construction

Petrova et al. (2018*c*) provide an extensive definition of KDD approaches according to the different types of building data (numeric data, semantic BIM data, geometric data, sensor data, etc.) and the knowledge discovery purpose. Due to the abundance of spatio-temporal data, the AEC industry can benefit from mining temporal data (time series) and spatial data. Shekhar et al. (2010) rightfully indicates that extracting interesting patterns and associations from such complex and multidimensional data with plenty of dependencies and spatio-temporal correlations is more difficult than mining traditional numeric and categorical data. In AEC, spatio-temporal data mining approaches can be valuable in cases where spatial data is augmented with time series data from diverse sensor networks in buildings or infrastructure.

Data mining applications for improvement of building performance and sustainable building design usually relate to energy use and demand prediction (Wang & Srinivasan 2017), prediction of occupant behavior (D'Oca & Hong 2014), fault detection for building systems (Cheng et al. 2016), improvement of building operation and optimal control strategies (Xiao & Fan 2014), as well as discovering and explaining energy use patterns (Miller et al. 2015). Other researchers have investigated the use of semantic data modelling, neural networks and data mining for building energy management (McGlinn et al. 2017). As can be seen from these examples, the use of KDD is usually related to the improvement of the building operation. Using such approaches to improve future building design processes have not been investigated in such detail. Examples of mining BIM data and simulation data for extraction of useful patterns in building design can be seen in Yarmohammadi et al. (2017).

Limitations in the application of Data Mining

"Classic" data mining techniques typically focus on isolated "silo" data. As stated by Lausch et al. (2015), in such cases, the conclusions remain limited and do not span interdisciplinary and complex data. Additionally, data selection and treatment resides in the hands of the analyst, who holds the responsibility for decision-making related to variable selection and data preparation to fit the needs of the mining algorithms. In case of incorrect decisions, the results can be influenced negatively, e.g. hidden patterns and novel knowledge may not be discovered or registered.

Therefore, Lausch et al. (2015) propose to mine data using linked data technologies. Such an approach allows opening silos and integrating data across disciplines, and provides an opportunity for analysis of interdisciplinary datasets. This broad overview can lead to insightful analyses, especially in a semantically rich domain such as AEC. Nevertheless, how these analyses are obtained is very different from the methods used in data mining, in the sense that the linked data realm is governed by queries and rules. These methods can be considered graph mining or matching techniques, and therefore potentially similar to pattern recognition. However, the types of graphs and patterns used in semantic queries and rules are very different from the patterns uncovered using data mining techniques, and both should not be perceived as identical.

Knowledge Graphs, Linked Data and the Semantic Web

Further to the evolutions in KDD, a lot of progress has been made in the formalization of knowledge using web technologies. From a web of documents, the World Wide Web has now evolved into a 'Web of Data' (Linked Open Data cloud) (Bizer et al. 2009). The term Linked Data was coined by Tim Berners-Lee in 2006[3] and has enabled worldwide publication of 5-star open data[4]. This implies defining data according to the Resource Description Framework (RDF)[5] data model and interlinking it with other RDF-based datasets available on the web. The Web of Data relies on ontologies so that data is typed and can

[3] http://www.w3.org/DesignIssues/LinkedData.html
[4] http://5stardata.info/
[5] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

easily be used in combination with query and rule languages such as SPARQL. Ontologies can be defined using RDFS and OWL[6] and give 'meaning' or 'semantics' to the data, thereby constituting the Semantic Web as conceived in Berners-Lee et al. (2001).

Due to their potential, linked data and semantic web technologies have received major attention in the AEC industry. A comprehensive overview of this topic can be found in Pauwels et al. (2017). Among the most notable initiatives is the early work on transforming the Industry Foundation Classes (IFC) into an OWL ontology (ifcOWL) (Pauwels & Terkaj 2016). The ifcOWL ontology was built to match the original EXPRESS schema as closely as possible, thus allowing a round-trip conversion process (lossless conversion). However, this has lead to a very big ontology, which resembles the IFC schema almost entirely, i.e., difficult to extend, complex, and not modular. This led to research initiatives aiming at ontologies for Linked Building Data, which do not rewind to IFC, yet cover similar ground.

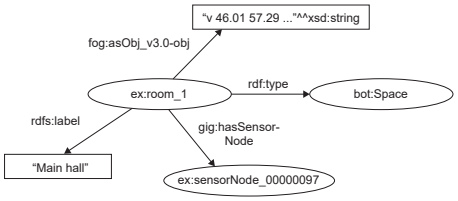

Figure 2: An example LBD graph.

At the moment, an ecosystem of smaller domain ontologies is available, each covering parts of what can also be handled with IFC (Fig. 2 and 3). A central Building Topology Ontology (BOT) Rasmussen et al. (2017) captures terms as 'Building', 'Site', 'Space', 'Element', etc. and aims for standardisation of these terms within the W3C LBD CG. Starting from BOT, alignments with various domain ontologies (Schneider 2017) can then be made. As a result, the industry can rely on a modular set of ontologies (Schneider et al. 2018), yet still have a stable standard at the core. Besides topology, other ontologies in the W3C LBD CG cover products, properties, and geometry McGlinn et al. (2019).

Semantic Data Mining

Standard data mining algorithms usually use statistical models on data to discover patterns and provide actionable insights. According to Lavrač et al. (2011), during that process, data is treated as meaningless numbers and attribute values. In other words, data by itself does not convey any semantic meaning and needs to be interpreted to present meaningful information, which is usually done by domain experts. Usually, such processes are associated with an abundance of raw data, but the underlying knowledge is scarce. Considering that KDD and data mining are knowledge-intensive processes, they can significantly benefit from enrichment by domain knowledge and the relations between objects. As further stated by Lavrač et al. (2011), that can be achieved by adding semantic descriptors (annotations) to the data and by the use of domain ontologies. This concept has caused a paradigm shift in data mining, expressed in a transition from mining the raw data to mining the knowledge directly. An overview of how semantic web technologies can be used in data mining and KDD is given in Ristoski & Paulheim (2016). Further studies on the concept of knowledge-based data mining have been performed by Barba-González et al. (2019).

The increased interest in the fusion of data mining and semantic has also highlighted the main technical challenges and opportunities that this union presents. For instance, classic data mining is powerful for extracting useful patterns and association rules from large traditional datasets. Yet, as Nebot & Berlanga (2012) state, the different nature of semantic data presents challenges, which cannot be tackled by traditional machine learning approaches, as they target mostly homogeneous data composed by transactions (sets of items). Since annotated data does not follow a rigid structure, instances, which are a part of the same class may still have a different structure. That causes one of the biggest challenges, i.e., structural heterogeneity. Together with the heterogeneity of data sources, this leads to the necessity of specifically dedicated approaches for pattern discovery in semantic data. This includes reasoning capabilities that allow inferring the implicit knowledge residing in the ontology itself (`subClassOf` relations, rules, inverse relations, etc.). For those reasons, several research efforts have engaged in defining the pathway towards effective association rule mining in knowledge bases (Barati et al. 2016, Galárraga et al. 2013).

Storing and processing sensor data

An important body of work in the semantic web domain, which is also of particular relevance in this paper, lies in the context of sensors and actuators. Sensor nodes are placed in precisely determined locations with a particular purpose of observation, thereby monitoring building use and performance in a real-time manner. This typically results in significantly large amounts of continuous data streams, often captured in data lakes. Such data can be used in RDF graphs (Semantic Sensor Networks), and thus be directly included as separate modules complementing the modular LBD cloud. Example ontologies that can be

[6]http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

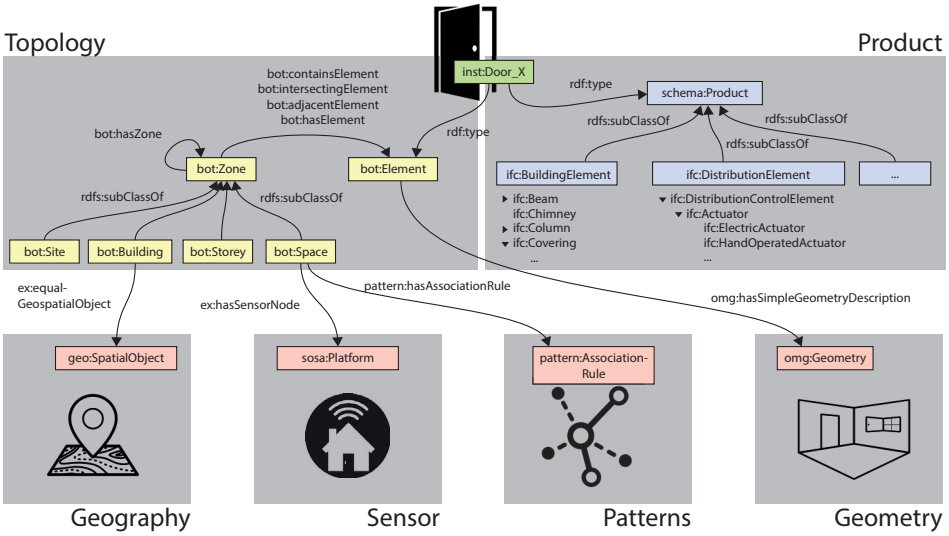

Figure 3: Conceptual overview of the modules and ontologies in a linked building data cloud, based on the work in the W3C LBD CG.

used for this purpose are SOSA[7], SSN[8] and SAREF[9].

Calbimonte et al. (2012) state that the heterogeneity of sensor data sources and environments is an important issue related to the realization of a connected sensor world. Monitored data is usually represented in different ways by different networks, and data models and schemas differ just as much. That leads to several compatibility and representation issues. To tackle those, research efforts propose various solutions such as semantic annotation of sensor data (Sheth et al. 2008), providing ontology-based access to data (Calbimonte et al. 2010), etc.

Storing the vast amount of data directly in the RDF graph typically leads to a "swollen" graph, and takes down query and reasoning performance. Hence, Petrova et al. (2018*a*,*b*) propose to maintain sensor data within their common non-RDF based data stores, yet link directly from the RDF graph to the web API providing access to the sensor data. When relying on web technologies for application development, these HTTP links can be consumed to give a custom, fast and on-demand access to the raw sensor data. However, several studies suggest that further opportunities may arise from using SPARQL queries with streaming extensions to access observations (Calbimonte et al. 2012). RDF stream processing may give an opportunity to publish and analyze real-time data streams while avoiding the "swollen" graph issue and still make sensor data a part of the LBD knowledge graph. Della Valle et al. (2009) state that achieving that would require moving from storing data and querying it on demand ("one-time semantics") to using continuous queries ("continuous semantics"). Barbieri et al. (2010) state that focus needs to be put on "stream reasoning", i.e., making sense of multiple real-time heterogeneous data streams. Llanes et al. (2016) define three main stages in the publication of RDF streams, i.e. conversion from sensor data streams to RDF streams, storing RDF streams, and linking them with other data sources. That requires the selection of relevant ontologies, defining the mapping language for conversion, selection of continuous query languages (e.g. Continuous SPARQL (C-SPARQL) and SPARQLstream (Barbieri et al. 2010), (Calbimonte et al. 2012)) and choosing other appropriate datasets to link to.

[7]http://www.w3.org/ns/sosa/
[8]http://www.w3.org/ns/ssn/
[9]http://ontology.tno.nl/saref/

Semantic Data Mining and Linked Data for a Recommender System in the AEC Industry

Conceptual framework

As previously stated, this article aims to outline the necessary steps for development of a system that relies on knowledge graphs to make recommendations for sustainable design decision support. Based on the state of the art, we conclude that in the implementation of the recommender system (1) knowledge graphs can be accessed

using semantically rich queries, (2) raw sensor data can be mined with traditional data mining techniques, (3) semantic data mining can be performed on the LBD graph, and (4) RDF graph mining techniques can also be used for pattern matching in combination with RDF stream processing.

Furthermore, a recommender system can rely on data sources both without and with explicitly embedded semantics. In the latter case, recommender systems rely directly on semantic analysis techniques (e.g. semantic data mining), thereby directly exploiting the semantics in the linked data graph. In the current context, in which the modular LBD graphs consist of both graph data (topology and product data) and non-graph data (geometry and sensor data), both traditional and semantic data mining can be used. On that note, it is important to distinguish between these two pattern discovery techniques and how they apply. Mining of raw sensor data implies discovery of performance patterns by the use of classic data mining methods. The knowledge interpretation is strictly related to obtaining understanding about the performance through the discovered patterns, not through the raw data. The RDF frequent pattern discovery, on the other hand, is data structure oriented and considers the graph predicates instead of data values.

Applying these techniques results in the conceptual system architecture in Fig. 4. The following sections explain this architecture in more detail, focusing on (1) how patterns are discovered and added to the graph, (2) how user profiles can be built and benefit from the system, including feedback, and (3) how recommendations can be generated. We present an example for RDF pattern discovery in a semantic data stream by implementing a method suggested by Belghaouti et al. (2016) and discuss its potential feasibility. Finally, we demonstrate an initial implementation of an linked data-based recommender system by applying the concept of Linked Data Semantic Distances proposed by Passant (2010).

Pattern discovery and representation

As a first step, data about existing buildings and design models is retrieved and transformed into linked data. We hereby suggest to rely on the overall LBD approach documented earlier in Petrova et al. (2018*a*,*b*). This process is displayed on the bottom right side in the system architecture diagram in Fig. 4. For describing sensors, the LBD graph can be enriched with sensor node instances and sensors, as can be seen in Fig. 2. Listing 1 lists all namespaces and prefixes used in the following examples.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix bot: <https://w3id.org/bot#> .
@prefix buildings: <https://www.example.com/data/buildings/> .
@prefix people: <https://www.example.com/data/people#> .
@prefix ls: <https://www.example.com/voc/linkset#> .
@prefix bmeta: <https://www.example.com/voc/buildingmetadata#>
    .
```

Listing 1: Namespaces and prefixes used in the following examples

As indicated in the state of the art section, including sensor measurements can then be done by pointing to an SQL store via a Web API or by including the sensor measurements explicitly in the graph. In this case, pattern discovery can be done using traditional data mining techniques, which work with batches of data and use the previously discussed predictive and/or descriptive models. As explained in Petrova et al. (2018*a*,*b*), the resulting performance patterns that have been mined from the sensor data values can also be stored directly in the graph.

Alternatively, it is possible to continuously convert the sensor data streams into RDF streams and perform semantic data mining on the resulting graph. Ideally, the RDF graph is first completed, which requires reasoning through the data and ontologies, and inferring all implicit data (e.g. `subclassOf` relations). To analyze how RDF stream processing would affect the recommendation concept, we can employ the method described by Belghaouti et al. (2016), who identify frequent RDF patterns in RDF streams by mapping the graphs to adjacency matrices based on the graph predicates. Using this method, one is able to construct bit vectors, which describe the graph structure. Each bit vector is constructed from the predicates in the graph. The graph in Fig. 2, for example, would lead to a bit vector (1111) that indicates the presence of each of the four predicates ('rdfs:label', 'gig:hasSensorNode', 'rdf:type', and 'fog:asObj- v3.0-obj'). All predicates and corresponding bit vector indices are recorded in a predicate hash table, which detects the patterns in the streams based on the bit vectors present in the graphs (e.g. 1111, 11101, 101, etc). Finally, a graph hash table is constructed, which records the frequency of occurrence of each bit vector. In this case, considering that all observations in the stream are modelled with the same predicates as in Fig. 2, only one pattern would be included in the graph hash table, even though very diverse observation measurements are present.

This has a big impact on pattern discovery, as the RDF frequent pattern discovery is data structure oriented and considers the graph predicates instead of data values, as opposed to traditional data mining techniques, which focus only on data values.

User profiling and feedback

User profiling is a required feature for a well-functioning user-centred recommender system. We have set up the

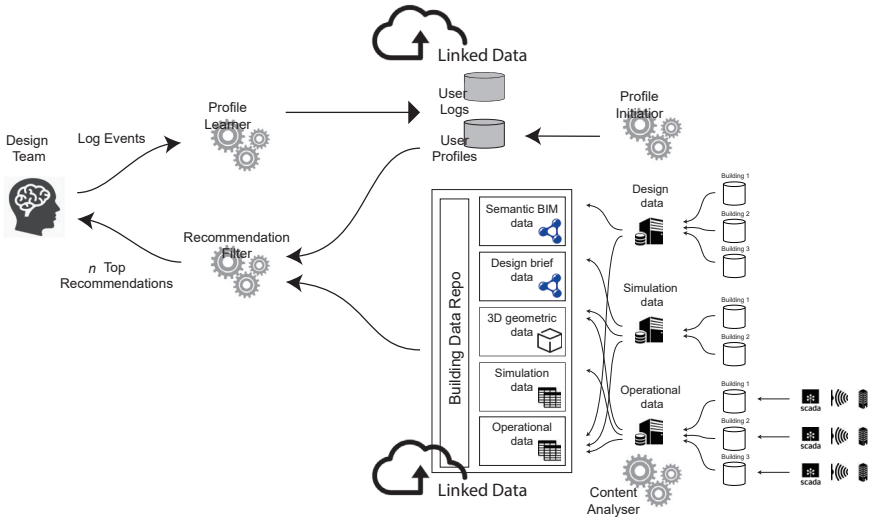

Figure 4: System Architecture for a linked data-based recommender system in the AEC industry.

profiling system in a way similar to the one proposed in Boratto et al. (2017) (top of Fig. 4). At user registration, a *Profile Initiator* component fills an RDF-based *User Profile Store*. These RDF-based profiles are built using the FOAF[10] ontology, and the result is an initial RDF graph identifying a user and its key metadata (Listing 2). The user is served recommendations through the *Recommendation Filter* component. All actions that the user takes in direct interaction with the recommender system are logged through a *Profile Learner* component. These actions thus serve as 'feedback' to the system, and they may come from a user clicking a 'like' button, a 'category' button, an 'annotation' button, or any other form of interaction (e.g. interactions identified by eye-tracking, clicking behavior, etc.). Listing all potential actions that a user takes and from which feedback is obtained, is out of scope for this paper. The Profile Learner component feeds back user profile data and user logs into the back-end of the recommendation system, which contains the *User Profile Store* and the *User Log Store*. In other words, the User Profile Store gets modified incrementally under the effect of the user interactions. The interactions of highest relevance are of course those related to recommendations, which are used by the end user in the project, especially if they respond directly to specific design requirements and/or performance targets.

```
people:EkaterinaPetrova
  a foaf:Person ;
  foaf:name "Ekaterina Petrova"^^xsd:string ;
  foaf:givenName "Ekaterina"^^xsd:string ;
  foaf:familyName "Petrova"^^xsd:string ;
  foaf:nick "epetrova"^^xsd:string .
```

Listing 2: People profile data

Feedback from user interaction goes into the User Logs and User Profiles, but the links between specific user profiles and relevant items in the *Building Data Store* are also kept, thereby aiming to enable a context-aware system. This means that we store links between user profiles and building identifiers in a separate RDF linkset (Listing 3). This linkset serves as a hash table with identifiers from user profiles and the building data repository. Note that Listing 3 only includes `ls:like` relations, but other, more specific relations could be used as well, depending on how user interaction and feedback is tracked.

```
people:EkaterinaPetrova
  ls:likes buildings:building_987d706d-877a-4b1d-80f6-6
    ee89d856319 ;
  ls:likes buildings:building_af41d889-f50c-456e
    -9625-96655150838d .
```

Listing 3: Linkset between buildings and people.

We have applied this principle to the Building Data Store, User Profile Store, and Linkset Store as follows. Through user interaction and knowledge discovery, implicit data is retrieved about the buildings in the building data repository. As such, the buildings can be enriched with metadata tags. The result is displayed for two example buildings in Listing 4. Whereas this example only includes

[10]http://xmlns.com/foaf/spec/

four simple metadata tags (`buildingType`, `designedBy`, `energyLabel`, `sustainabilityCertificate`), many more metadata tags can be used, e.g. category, occupancy data, mined performance patterns, design requirements, energy source, etc. These metadata can be used to form categories of design references, to compose queries in the database, to sort search results in a certain dimension, etc.

```
buildings:building_00dd6c87-6a6e-f482-7490-e6613659708a
  a bot:Building ;
  bmeta:buildingType bmeta:theater ;
  bmeta:designedBy people:architectX ;
  bmeta:energyLabel bmeta:A ;
  bmeta:sustainabilityCertificate bmeta:LEEDPlatinum .

buildings:building_2e0dcc1c-b981-4c47-adb4-2b9887f10481
  a bot:Building ;
  bmeta:buildingType bmeta:theater ;
  bmeta:designedBy people:architectY ;
  bmeta:energyLabel bmeta:A ;
  bmeta:sustainabilityCertificate bmeta:DGNBGold .
```

Listing 4: Example building data in TTL format.

In summary, the system holds three RDF-based data stores (besides the User Log Store): the User Profile Store, the Building Data Store, and the Linkset Store. It is now possible for an end user to query each of these stores for relevant data. For example, an end user may fire a query for all buildings of a particular type, category and/or with a specific energy label (Listing 5). In this case the `bmeta` tags are used in the query. Of course, it is also possible to include user preference (Linkset Store) or user profile (User Profile Store) data in the queries. The returned results can be displayed to an end user, who is then able to sort the results using the available attributes and categories.

```
SELECT *
WHERE {
    ?b a bot:Building .
  ?b bmeta:buildingType bmeta:theater ;
    ?b bmeta:energyLabel bmeta:A .
}
```

Listing 5: Query for buildings of a particular building type.

Generating recommendations

As stated in the state of the art section, recommender systems often rely on the computation of the semantic distance between concepts, or in other words, the semantic relatedness between two resources. Instead of limiting only to queries that can be sent from within the end user environment (previous section), our recommender system should also make suggestions of buildings that are semantically close to, for example, a building that is considered to be most relevant to an end user at some point in time. Such buildings are the generated recommendations.

A set of measures were proposed in Passant (2010) to represent the 'Linked Data Semantic Distance' (*LDSD*) between two concepts (values between 0 and 1). This includes Direct, Indirect, and Combined Semantic Distance ($LDSD_d$, $LDSD_i$, $LDSD_c$), each either weighted or not. These semantic distances are used in recommender systems to find out what else users may like based on their user profile, search behavior, favorites, etc. The smaller the semantic distance between two related concepts, the higher the related concept is ranked in the set of n top related concepts or recommendations.

The semantic distance can be computed using all outgoing and incoming links of two concepts. For example, two different buildings might both be of type `theater`, which connects them to the same node for the `bmeta:buildingType` predicate, and makes them semantically closer. Determination of LDSD for recommendations starts as soon as an end user clicks a building from a result set that was previously returned with a simple query. In other words, the Recommendation Filter component is set up to look for 'bot:Building' objects that are semantically close to each other. The calculation hereby relies on all incoming and outgoing links for specific buildings, which are linked in the Building Data Store and the Linkset Store. Essentially, the simple indirect distance as a matrix between one building and all related buildings is calculated (Passant 2010).

This is illustrated in a simple example in Table 1, which shows the semantic distances for one of the buildings in the Building Data Store. As the `bot:Building` tag is present for all concepts, it is disregarded. Of course, in this limited example with 6 buildings and 3 relations (`buildingType`, `designedBy`, `energyLabel`), values are quite far apart (1/3, 1/2, 1, or 0), because only three links are considered. In an actual Building Data Store, semantic distances are much more interesting and diverse.

For each of the retrieved buildings, any available data can be displayed. This may of course also include sensor measurement data and patterns found in those data, depending on the implementation method and storage system chosen for such data. Also metadata and user data can be displayed, in support of the end user. Of course, this data needs to be displayed in an appropriate end-user interface, which is out of scope here.

Challenges and limitations

In terms of effectiveness of the proposed system, potential challenges may need to be overcome. Generally, they can be related to, for instance, the user behaviour or the method that the recommendations are based on. Besides the knowledge base, the user and their preferences play an important role in a recommender system. Important to consider are changes over time in user profiling and preferences, which need to be taken into account continuously. Furthermore, end users may have similar profiles,

Table 1: Simple indirect semantic distances computed for https://www.example.com/data/buildings/building_00dd6c87-6a6e-f482-7490-e6613659708a.

Building	Cio	Cii	LDSD
https://www.example.com/data/buildings/building_2e0dcc1c-b981-4c47-adb4-2b9887f10481	2	0	0.3333
https://www.example.com/data/buildings/building_987d706d-877a-4b1d-80f6-6ee89d856319	1	0	0.5
https://www.example.com/data/buildings/building_43576e80-cf8c-11e1-8000-68a3c4d40f59	1	0	0.5
https://www.example.com/data/buildings/building_af41d889-f50c-456e-9625-96655150838d	0	0	1.0
https://www.example.com/data/buildings/building_aac3427f-eeb0-460c-ba47-14fd44c8be74	0	0	1.0

but different behaviour and preferences depending on the context, so they cannot be generalised. These phenomena can clearly affect the accuracy of a recommendation system, as the wrong user preferences may be considered by the system. Anomalous behaviour such as rejection or disliking of particular recommendations also needs to be analysed and factored in.

Another limitation may stem from the fact that despite being efficient, the LDSD approach only computes the semantic distance between two resources that are directly or indirectly linked through an intermediate resource. Therefore, enhanced LDSD algorithms may need to be used to expand the range beyond the two links distance. Also, in the current system, we only consider semantic distances between buildings. Other semantic distances may be used as well, to configure and refine the recommender system.

Conclusions

Recent years show a rapid increase in technology uptake, aiming to reduce the negative environmental contribution from the built environment. In our research, we particularly look into mitigating these problems at the source, by informing the design team with evidence-based feedback stemming from the existing building stock through a recommender system. Research on recommender systems has a relatively long history, but is seldom actively implemented in the AEC industry. In this paper, we attempt to overcome this challenge by the use of data mining and linked data technologies. In particular, this paper includes an extensive state of the art review in the areas of KDD, semantic web technologies, data stream processing and recommender systems. Furthermore, we investigate how to make sensor data streams efficiently available to the end user in addition to discovered knowledge. This may be achieved through semantic sensor data modelling, web API connections, and/or sensor data stream processing. Together with the broad review, we outline the necessary steps towards implementing a linked data-based recommender system, thereby drawing on those techniques that show most promising value from the literature review. The software architecture of this recommender system consists of triple stores, mechanisms for feedback handling, mechanisms for recommendations, mechanisms for data mining, and an interactive user interface. Future work should focus on further implementation in practice. This will include a challenge of hidden knowledge discovery, namely, how can the metadata tags be inferred in the most intelligent and informative manner. Furthermore, the way in which sensor data are combined with semantic data (explicit semantic modelling, Web APIs, stream reasoning), so that they can be used effectively in recommendation filtering, needs to be further investigated.

References

Barati, M., Bai, Q. & Liu, Q. (2016), SWARM: An Approach for Mining Semantic Association Rules from Semantic Web Data, *in* 'PRICAI 2016: Trends in Artificial Intelligence', Springer International Publishing, Cham, pp. 30–43.

Barba-González, C., García-Nieto, J., del Mar Roldán-García, M., Navas-Delgado, I., Nebro, A. J. & Aldana-Montes, J. F. (2019), 'BIGOWL: Knowledge centered Big Data analytics', *Expert Systems with Applications* **115**, 543–556.

Barbieri, D. F., Braga, D., Ceri, S., Della Valle, E. & Grossniklaus, M. (2010), 'C-SPARQL: A continuous query language for RDF data streams', *International Journal of Semantic Computing* **4**(1), 3–25.

Belghaouti, F., Bouzeghoub, A., Kazi-Aoul, Z. & Chiky, R. (2016), Fregrapad: Frequent rdf graph patterns detection for semantic data streams, *in* '2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)', pp. 1–9.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001), 'The Semantic Web', *Scientific American* **284**(5), 34–43.

Bizer, C., Heath, T. & Berners-Lee, T. (2009), 'Linked data - the story so far', *International Journal on Semantic Web and Information Systems* **5**(3), 122.

Boratto, L., Carta, S., Fenu, G. & Saia, R. (2017), 'Semantics-aware content-based recommender systems: Design and architecture guidelines', *Neurocomputing* **254**, 79–85.

Borrmann, A., König, M., Koch, C. & Beetz, J. (2018), *Building Information Modeling: Technology Foundations and Industry Practice*, 1 edn, Springer.

Calbimonte, J.-P., Corcho, O. & Gray, A. J. G. (2010), Enabling ontology-based access to streaming data sources, *in* 'The Semantic Web – ISWC 2010', Springer, Berlin, Heidelberg, pp. 96–111.

Calbimonte, J.-P., Jeung, H., Corcho, O. & Aberer, K. (2012), 'Enabling query technologies for the semantic sensor web', *International Journal on Semantic Web and Information Systems* **8**(1), 43–63.

Cheng, Z., Zhao, Q., Wang, F., Chen, Z., Jiang, Y. & Li, Y. (2016), Case studies of fault diagnosis and energy saving in buildings using data mining techniques, *in* 'Proceedings of IEEE international conference on automation science and engineering', IEEE, Fort Worth, TX, pp. 645–651.

Corry, E., Pauwels, P., Hu, S., Keane, M. & O'Donnell, J. (2015), 'A performance assessment ontology for the environmental and energy management of buildings', *Automation in Construction* **57**(September), 249–259.

de Gemmis, M., Lops, P., Musto, C., Narducci, F. & Semeraro, G. (2015), Semantics-aware content-based recommender systems, *in* 'Recommender Systems Handbook', Springer, p. 119159.

de Wilde, P. (2014), 'The gap between predicted and measured energy performance of buildings: A framework for investigation', *Automation in Construction* **41**, 40 – 49.

Della Valle, E., Ceri, S., van Harmelen, F. & Fensel, D. (2009), 'It's a streaming world! reasoning upon rapidly changing information', *IEEE Intelligent Systems* **24**(6), 83–89.

D'Oca, S. & Hong, T. (2014), 'A data-mining approach to discover patterns of window opening and closing behavior in offices', *Building and Environment* **82**, 726–739.

El-Diraby, T. E. (2013), 'Domain ontology for construction knowledge', *Journal of Construction Engineering and Management* **139**(7), 768–784.

Fan, C., Xiao, F., Madsen, H. & Wang, D. (2015), 'Temporal knowledge discovery in big bas data for building energy management', *Energy and Buildings* **109**, 75–89.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI Magazine* **17**(3), 37–54.

Galárraga, L. A., Teflioudi, C., Hose, K. & Suchanek, F. (2013), Amie: Association rule mining under incomplete evidence in ontological knowledge bases, *in* 'Proceedings of the 22nd International Conference on World Wide Web (WWW2013)', ACM, pp. 413–422.

Han, J., Kamber, M. & Pei, J. (2012), *Data mining concepts and techniques*, Morgan Kaufmann, Waltham, US.

Hand, D., Mannila, H. & Smyth, P. (2011), *Principles of Data Mining*, MIT Press.

Heitmann, B. & Hayes, C. (2010), Using linked data to build open, collaborative recommender systems, *in* 'AAAI spring symposium 2010: Linked data meets artificial intelligence', pp. 76–81.

Heylighen, A., Martin, M. & Cavallin, H. (2007), 'Building stories revisited: Unlocking the knowledge capital of architectural practice', *Architectural Engineering and Design Management* **3**(1), 65–74.

Lausch, A., Schmidt, A. & Tischendorf, L. (2015), 'Data mining and linked open data new perspectives for data analysis in environmental research', *Ecological Modelling* **295**, 5–17.

Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I. & Novak, P. K. (2011), Using Ontologies in Semantic Data Mining with SEGS and g-SEGS, *in* 'Discovery Science', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 165–178.

Llanes, K., Casanova, M. & Lemus, N. (2016), 'From sensor data streams to linked streaming data: A survey of main approaches', *Journal of Information and Data Management* **7**(2), 130–140.

McGlinn, K., Wagner, A., Bonsma, P., McNerney, L. & O'Sullivan, D. (2019), 'Interlinking geospatial and building geometry with existing and developing standards on the web', *Automation in Construction* . in press.

McGlinn, K., Yuce, B., Wicaksono, H., Howell, S. & Rezgui, Y. (2017), 'Usability evaluation of a web-based tool for supporting holistic building energy management', *Automation in Construction* **84**, 154–165.

Middleton, S. E., Shadbolt, N. R. & De Roure, D. C. (2004), 'Ontological user profiling in recommender systems', *ACM Transactions on Information Systems (TOIS)* **22**(1), 54–87.

Miller, C., Nagy, Z. & Schlueter, A. (2015), 'Automated daily pattern filtering of measured building performance data', *Automation in Construction* **49**, 1–17.

Miller, C., Nagy, Z. & Schlueter, A. (2018), 'A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings', *Renewable and Sustainable Energy Reviews* **81**, 1365 – 1377.

Molina-Solana, M., Ros, M., Ruiz, M., Gómez-Romero, J. & Martin-Bautista, M. (2017), 'Data science for building energy management: A review', *Renewable and Sustainable Energy Reviews* **70**, 598–609.

Musto, C., Basile, P., Lops, P., de Gemmis, M. & Semeraro, G. (2017), 'Introducing linked open data in graph-based recommender systems', *Information Processing and Management* **53**, 405–435.

Nebot, V. & Berlanga, R. (2012), 'Finding association rules in semantic web data', *Knowledge-Based Systems* **25**(1), 51–62.

Oliveira, J., Delgado, C. & Assaife, A. (2017), 'A recommendation approach for consuming linked open data', *Expert Systems With Applications* **72**, 407–420.

ODonnell, J., Corry, E., Hasan, S., Keane, M. & Curry, E. (2013), 'Building performance optimization using cross-domain scenario modeling, linked data, and complex event processing', *Building and Environment* **62**, 102 – 111.

Passant, A. (2010), Measuring semantic distance on linking data and using it for resources recommendations, *in* 'AAAI spring symposium 2010: Linked data meets artificial intelligence', pp. 93–98.

Pauwels, P. (2014), 'Supporting decision-making in the building life-cycle using linked building data', *Buildings* **3**, 549–579. DOI: 10.3390/buildings4030549.

Pauwels, P. & Terkaj, W. (2016), 'EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology', *Automation in Construction* **63**, 100–133.

Pauwels, P., Zhang, S. & Lee, Y.-C. (2017), 'Semantic web technologies in aec industry: a literature review', *Automation in Construction* **73**, 145–165.

Petrova, E., Pauwels, P., Svidt, K. & Jensen, R. (2018*a*), From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches, *in* 'Proceedings of the 11th European Conference on Product and Process Modelling (ECPPM)', pp. 391–398.

Petrova, E., Pauwels, P., Svidt, K. & Jensen, R. (2018*b*), In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data, *in* 'Proceedings of the CIB W78 Conference', pp. 19–26.

Petrova, E., Pauwels, P., Svidt, K. & Jensen, R. (2018*c*), 'Towards data-driven holistic sustainable design: A decision support framework relying on knowledge discovery in real-time building performance data and disparate project data repositories', *Architectural Engineering and Design Management* pp. 1–23.

Rasmussen, M. H., Pauwels, P., Hviid, C. A. & Karlshøj, J. (2017), Proposing a central AEC ontology that allows for domain specific extensions, *in* 'Proceedings of the Joint Conference on Computing in Construction'.

Ristoski, P. & Paulheim, H. (2016), 'Semantic web in data mining and knowledge discovery: A comprehensive survey', *Web Semantics: Science, Services and Agents on the World Wide Web* **36**, 1–22.

Sacks, R., Lee, C. M. E. G. & Teicholz, P. (2018), *BIM handbook: a guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators*, 3 edn, John Wiley & Sons, Hoboken, NJ, USA.

Schneider, G. F. (2017), Towards aligning domain ontologies with the Building Topology Ontology, *in* '5th Linked Data in Architecture and Construction Workshop'.

Schneider, G. F., Rasmussen, M. H., Bonsma, P., Oraskari, J. & Pauwels, P. (2018), Linked Building Data for Modular Building Information Modelling of a Smart Home, *in* 'eWork and eBusiness in Architecture, Engineering and Construction (ECPPM 2018)', pp. 407–414.

Shekhar, S., Zhang, P. & Huang, Y. (2010), Spatial data mining, *in* 'Data Mining and Knowledge Discovery Handbook 837-854.', Springer.

Sheth, A., Henson, C. & Sahoo, S. (2008), 'Semantic sensor web', *IEEE Internet Computing* **12**(4), 78–83.

Wang, Z. & Srinivasan, R. (2017), 'A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models', *Renewable and Sustainable Energy Reviews* **75**, 796–808.

Xiao, F. & Fan, C. (2014), 'Data mining in building automation system for improving building operational performance', *Energy and Buildings* **75**, 109–118.

Yarmohammadi, S., Pourabolghasem, R. & Castro-Lacouture, D. (2017), 'Mining implicit 3d modeling patterns from unstructured temporal bim log text data', *Automation in Construction* **81**, 17–24.